



Machine Learning for Environmental Applications

Hanson Center for Space Sciences
University of Texas at Dallas
Prof. David J. Lary



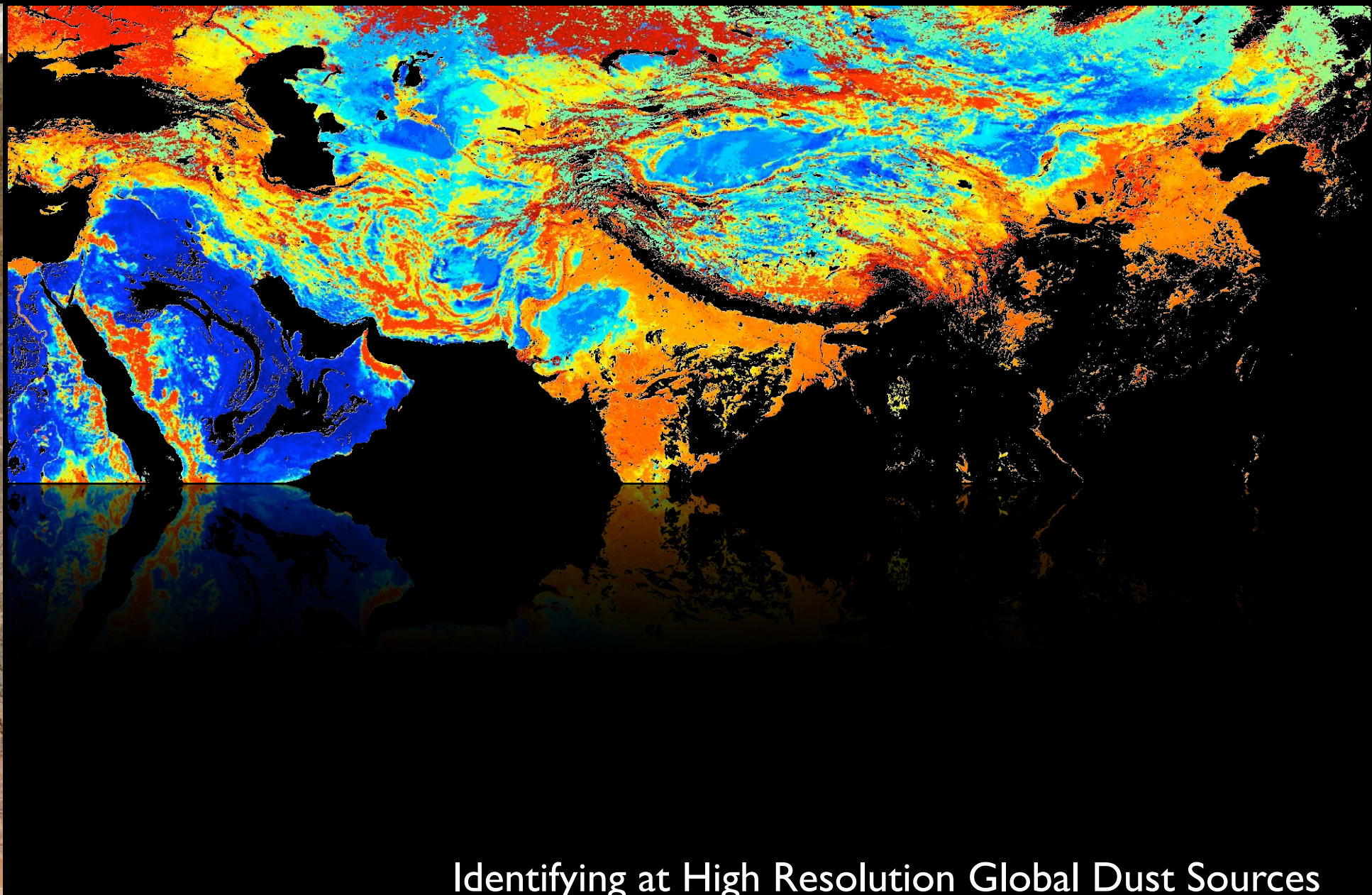
Detecting Dust Sources



Identifying at High Resolution Global Dust Sources

Global Distribution of Airborne Particulates

Detecting Dust Sources



Identifying at High Resolution Global Dust Sources

Global Distribution of Airborne Particulates

A Haboob (Arabic: هَبُوب “strong wind”, or “blowing furiously.”)



Midday

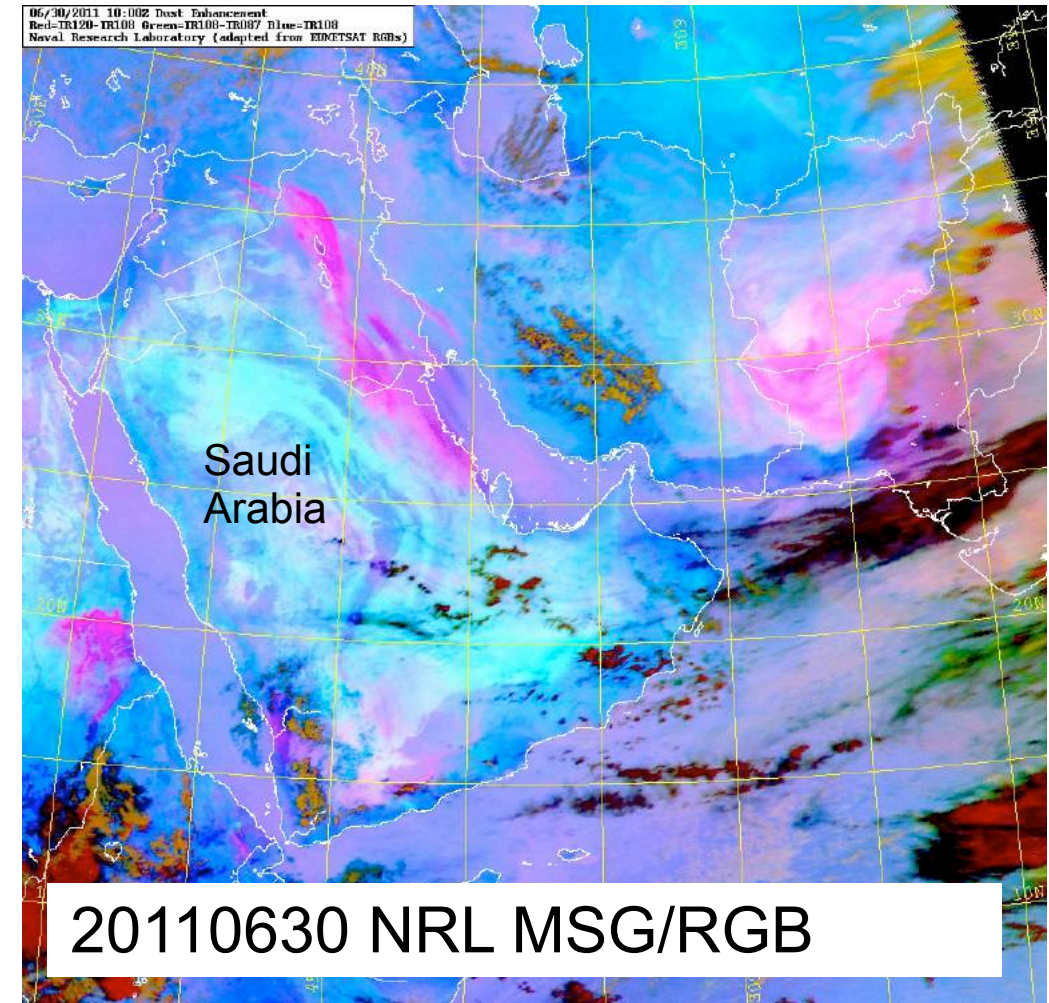
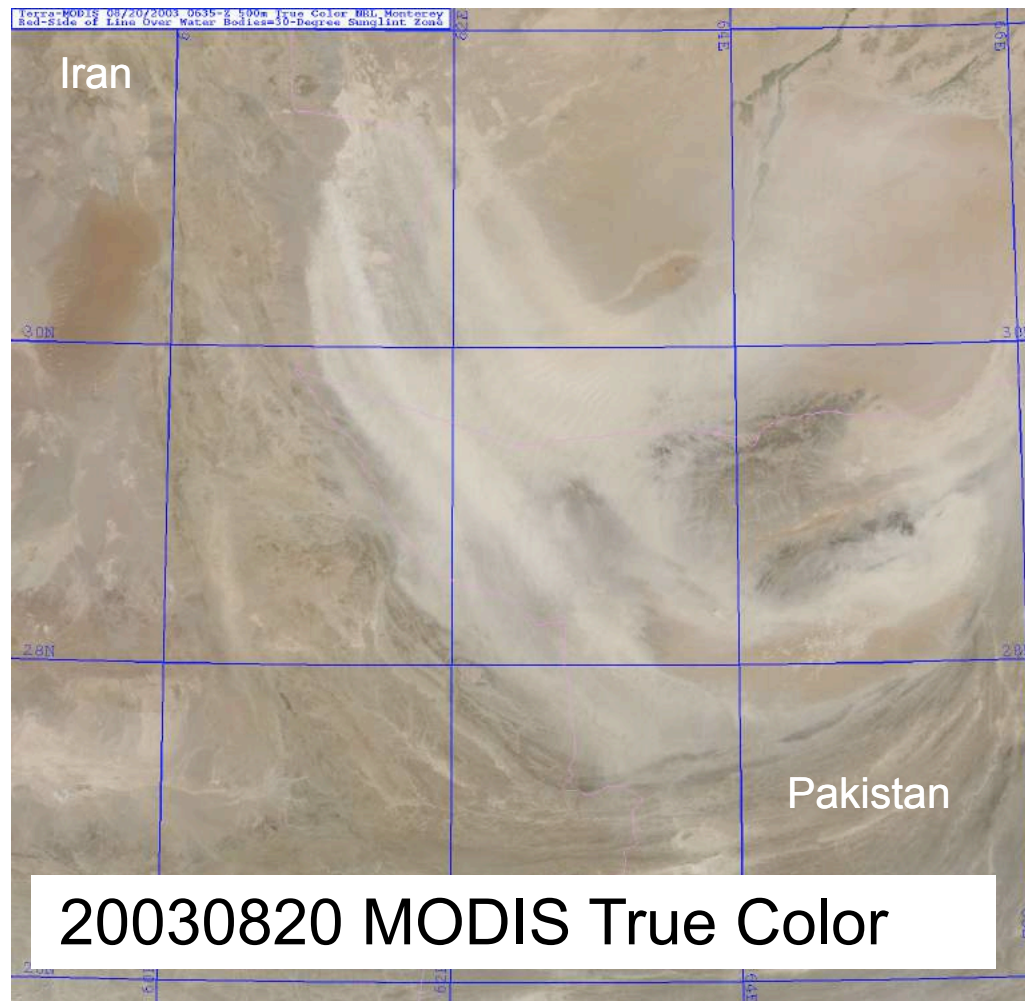
A Haboob (Arabic: هَبُوب “strong wind”, or “blowing furiously.”)



Midday



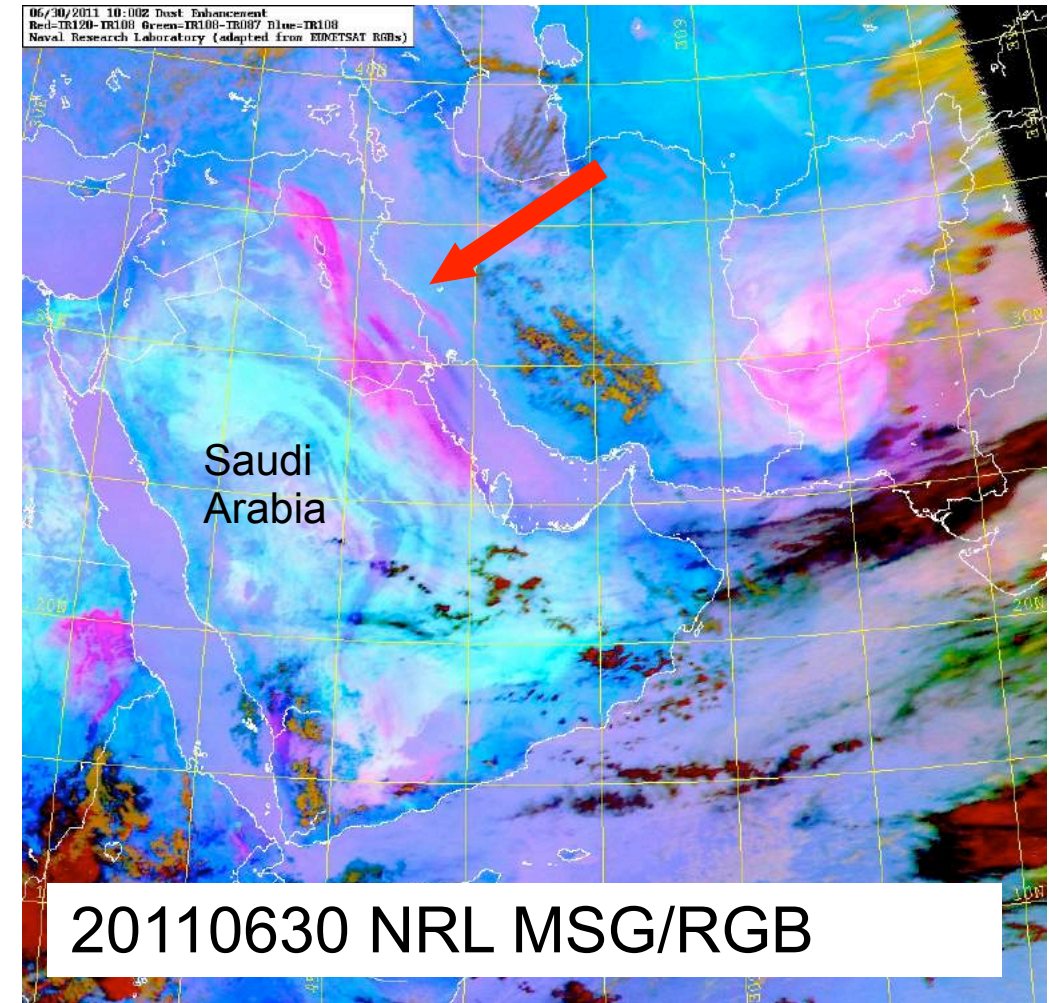
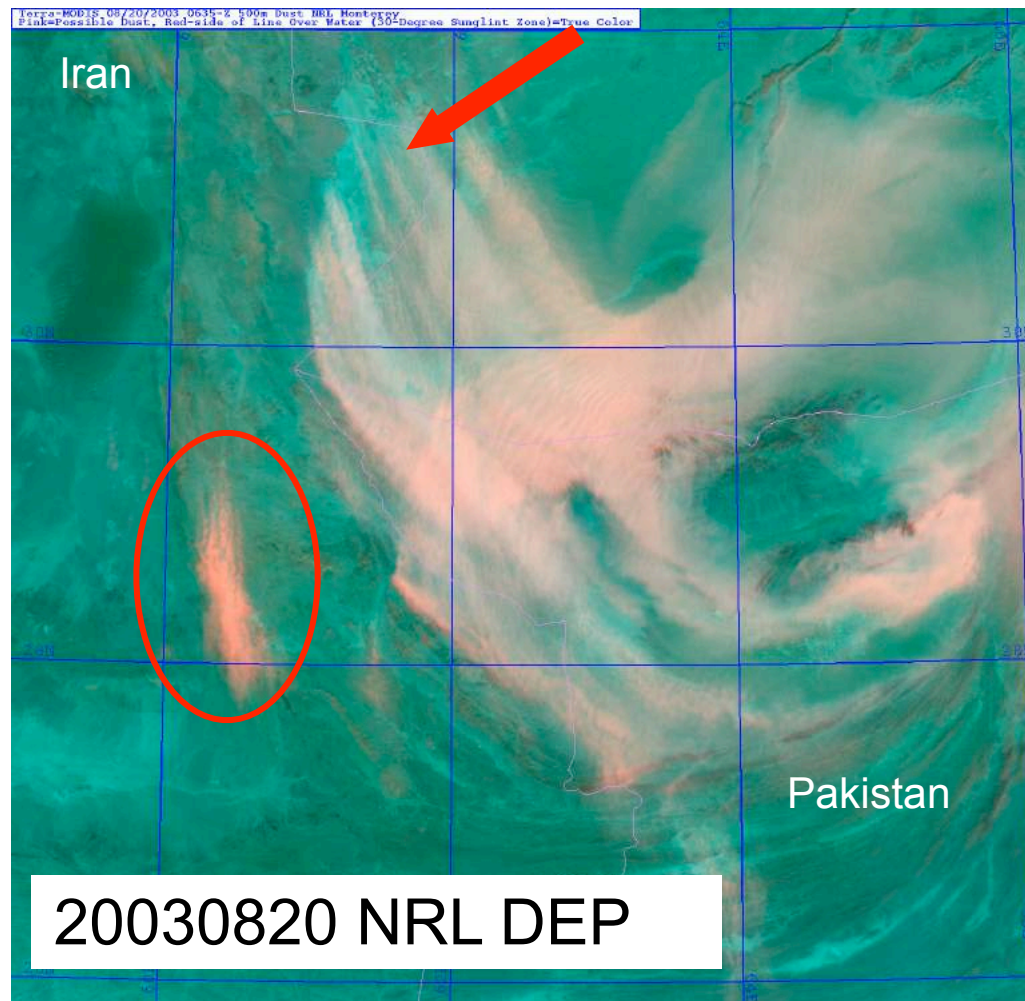
NRL High-resolution Dust Source Database



Approach and Methodology

- **10 years of DEP (2 yr MSG/RGB) imagery**
- COAMPS 10 m wind overlays
- Surface weather plots
- **ENVI (Gis-like software)**
- NGDC topographical 10°X10° tiles
- **Overlay 0.25° grid or use Google Earth (GE)**
- **Dust source area entered into database** (cursor location tool = 1km precision)
- Cross-correlate land and water features using maps, atlases, Landsat images (detailed topographic, geographic, and geomorphic information, **GE**)
- Technical and governmental reports

NRL High-resolution Dust Source Database



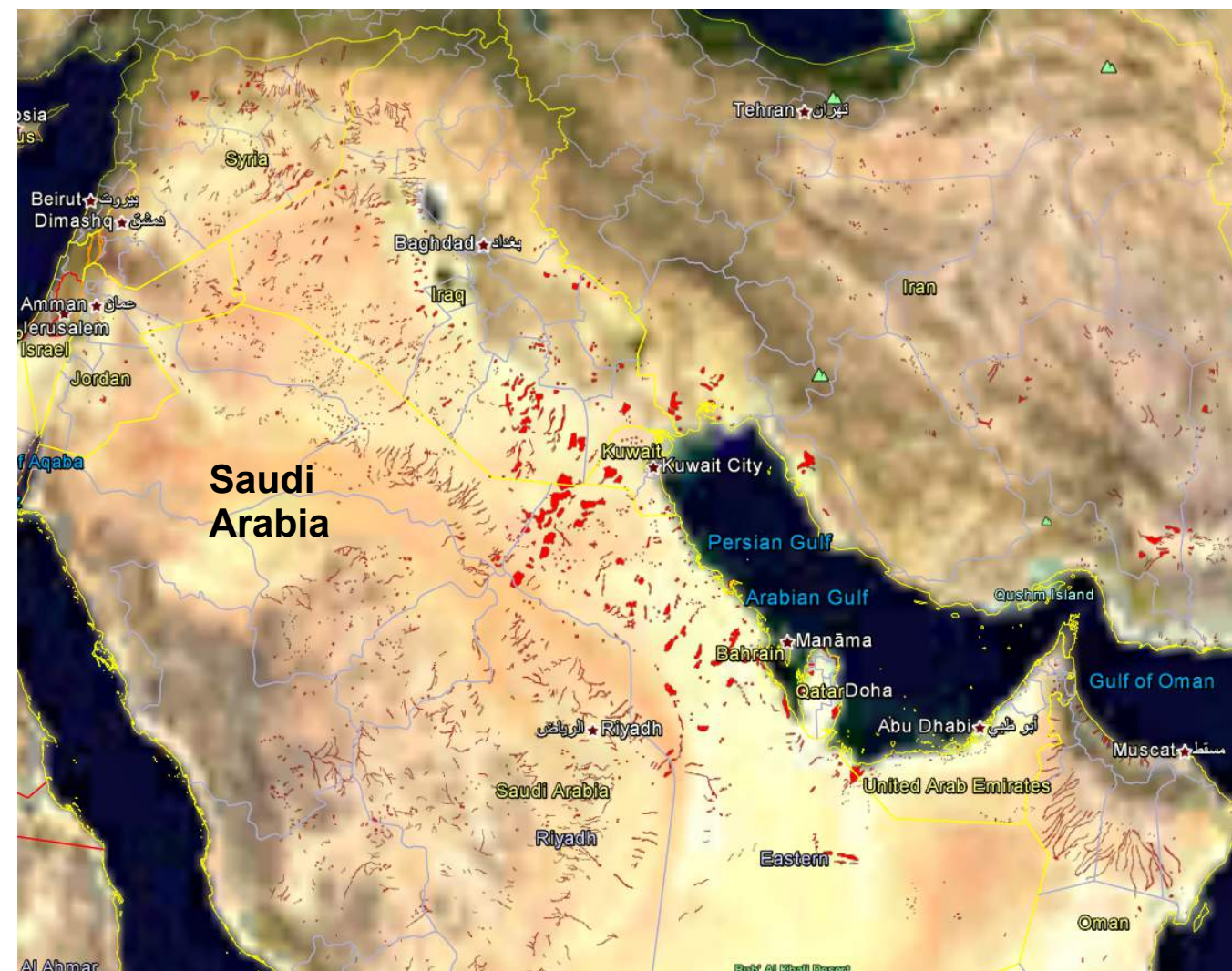
Approach and Methodology

- **10 years of DEP (2 yr MSG/RGB) imagery**
- COAMPS 10 m wind overlays
- Surface weather plots
- **ENVI (Gis-like software)**
- NGDC topographical 10°X10° tiles
- **Overlay 0.25° grid or use Google Earth (GE)**
- **Dust source area entered into database**
(cursor location tool = 1km precision)
- Cross-correlate land and water features using maps, atlases, Landsat images (detailed topographic, geographic, and geomorphic information, **GE**)
- Technical and governmental reports

NRL High-resolution Dust Source Database

Solid red and purple shapes identify dust source areas located using DEP and MSG.

SW Asia DSD



East Asia DSD

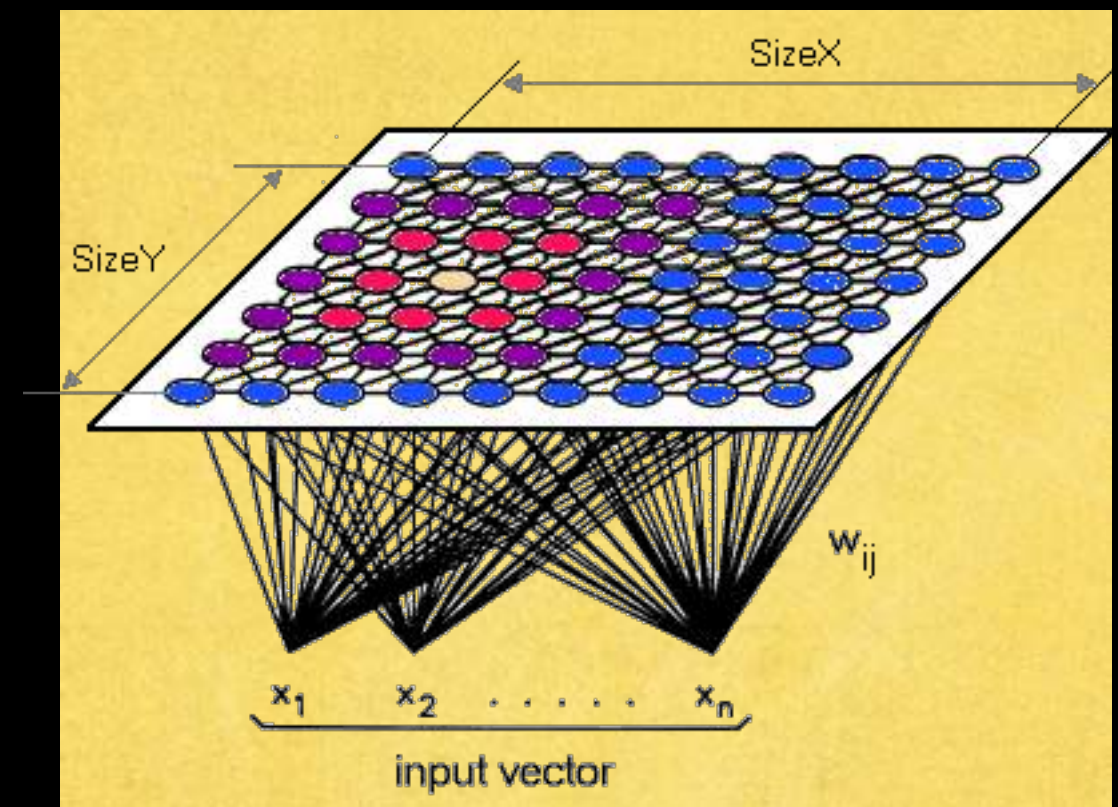
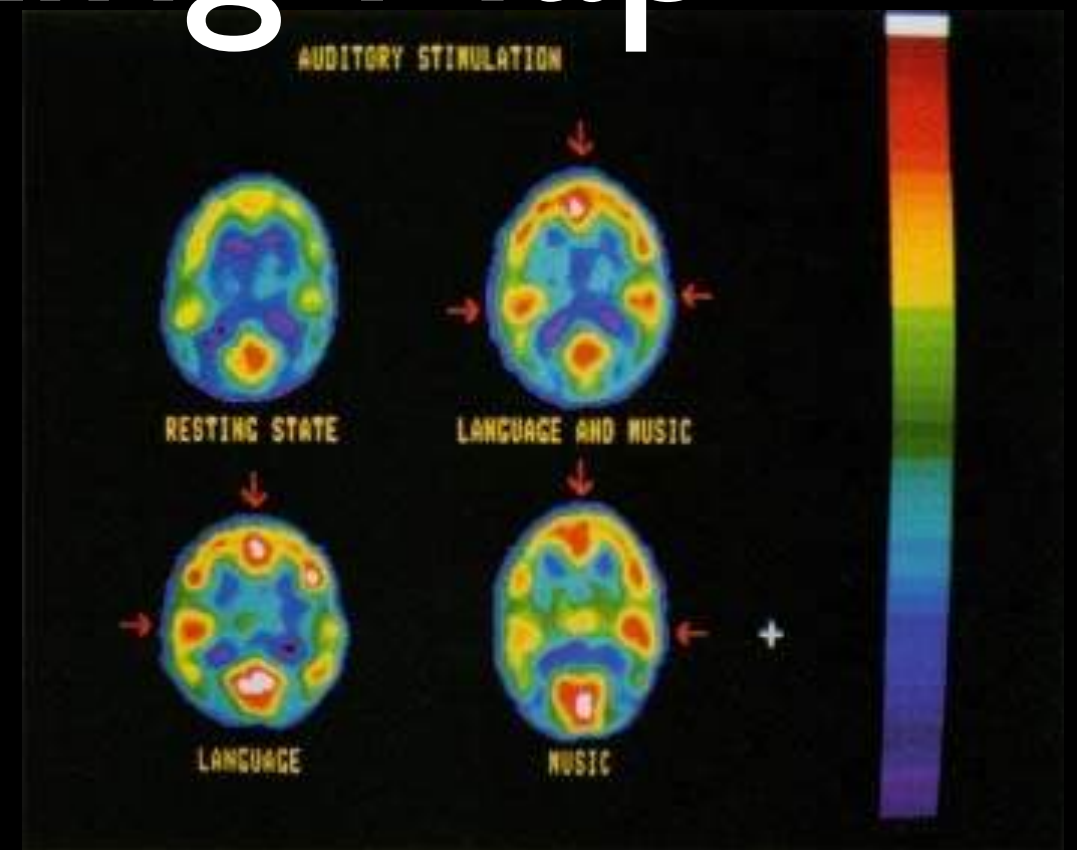


Self-Organizing Map

SOMs reduce dimensionality by producing a map that objectively plots the similarities of the data by grouping similar data items together.

SOMs learn to classify input vectors according to how they are grouped in the input space.

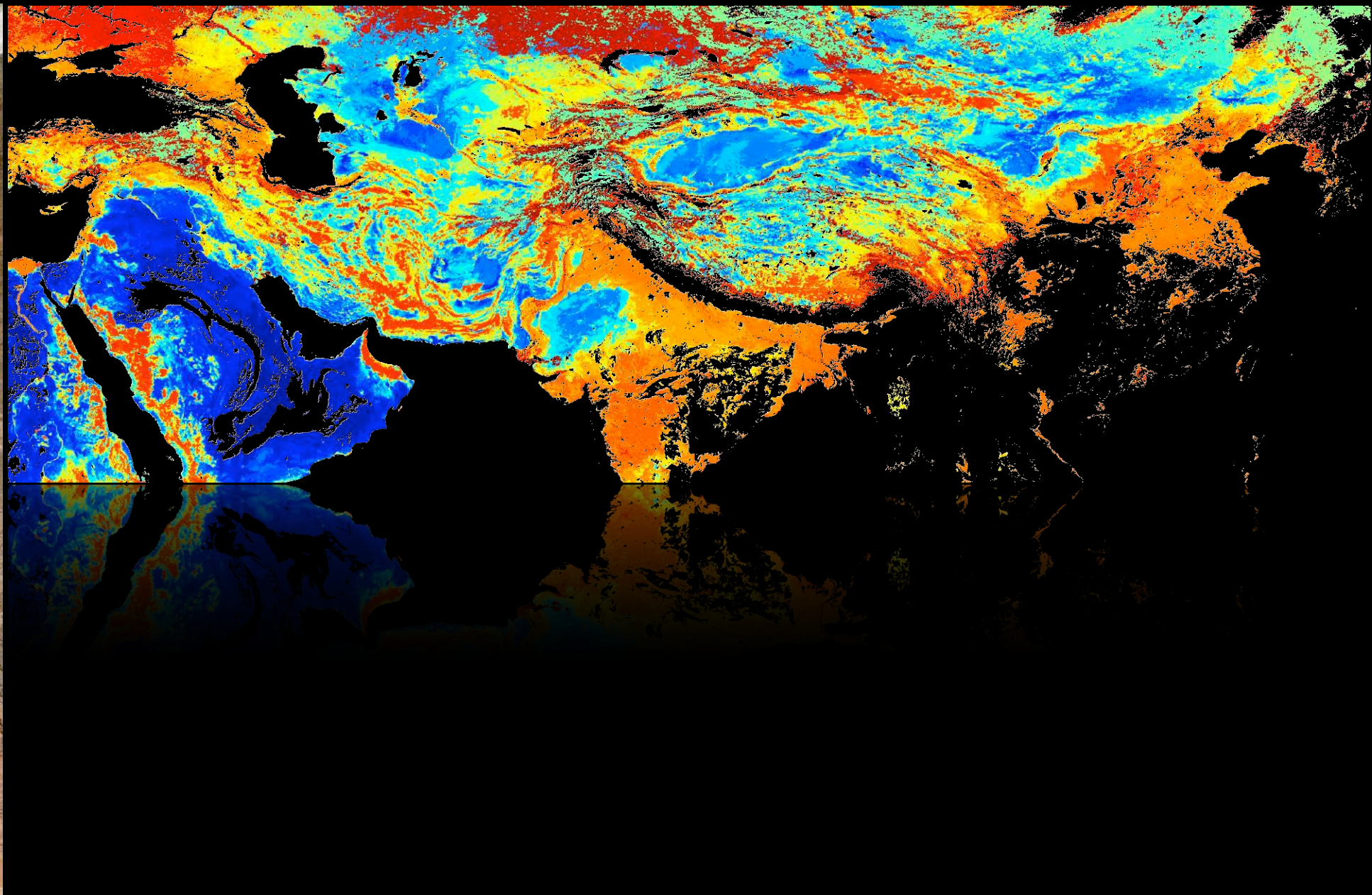
SOMs learn both the distribution and topology of the input vectors they are trained on. This approach allows SOMs to accomplish two things, reduce dimensions and display similarities.



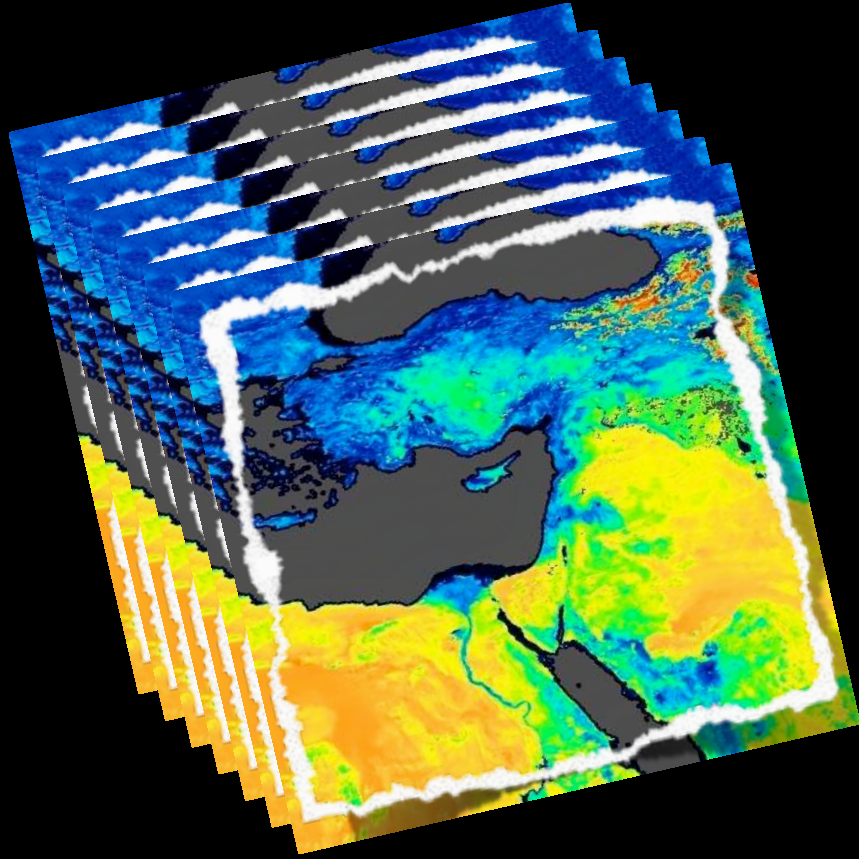
Detecting Dust Sources



Detecting Dust Sources

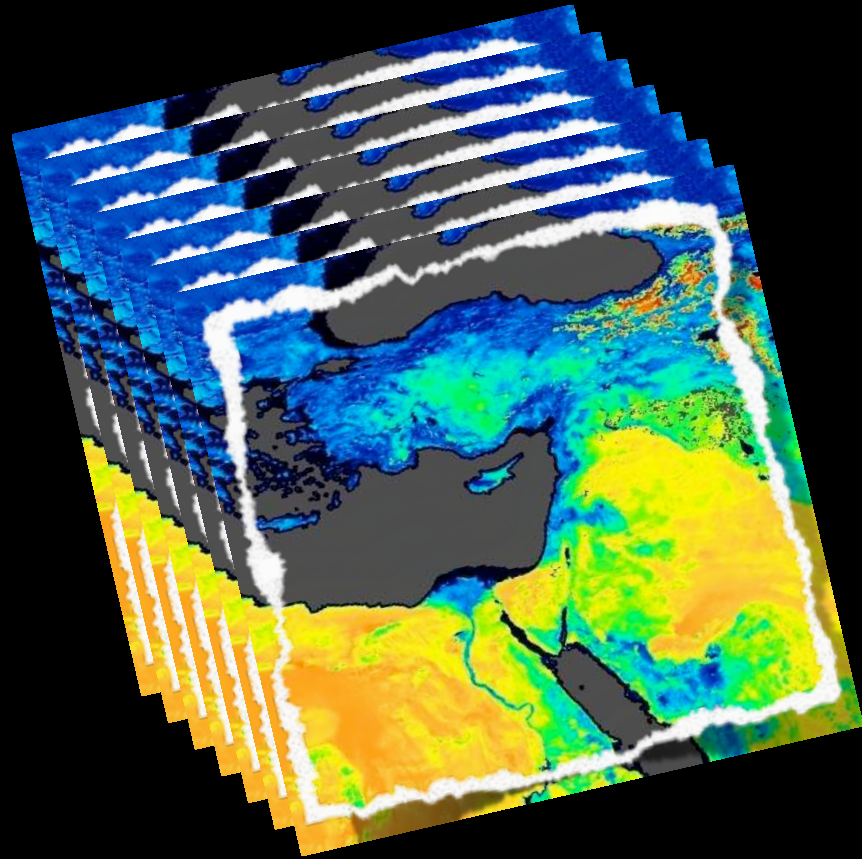


Self Organizing Map Classification

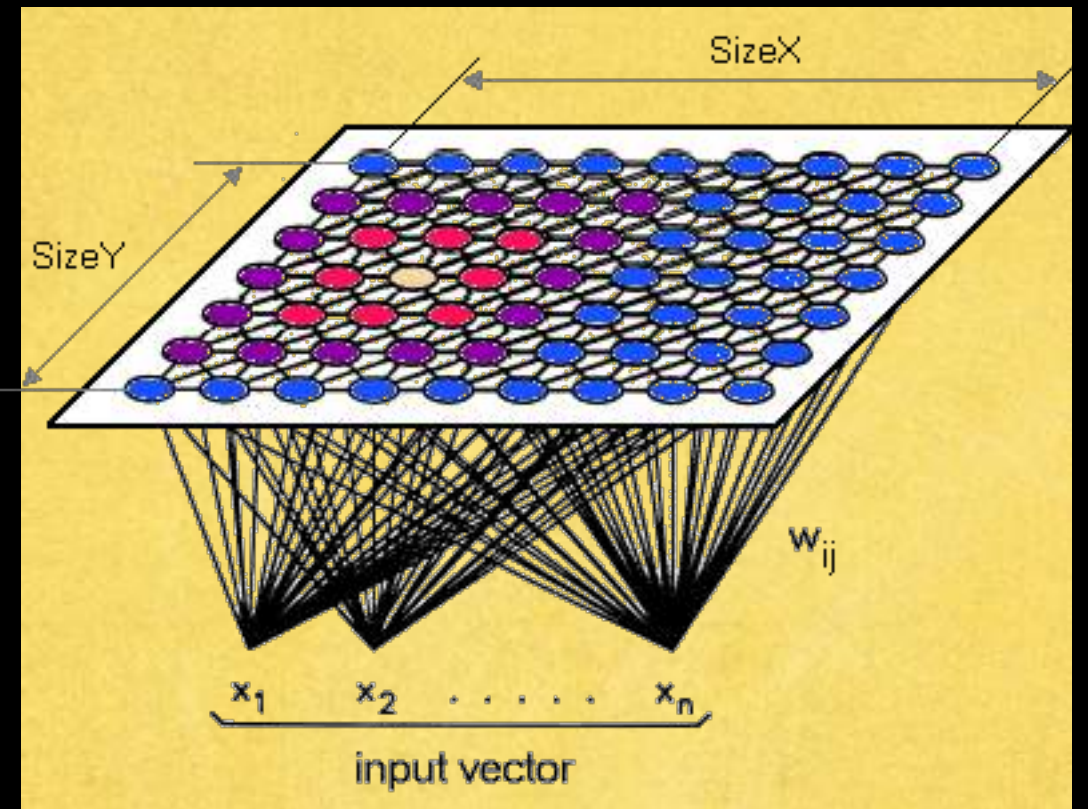


7 Bands
MODIS MCD43C3
bihemispherical reflectance

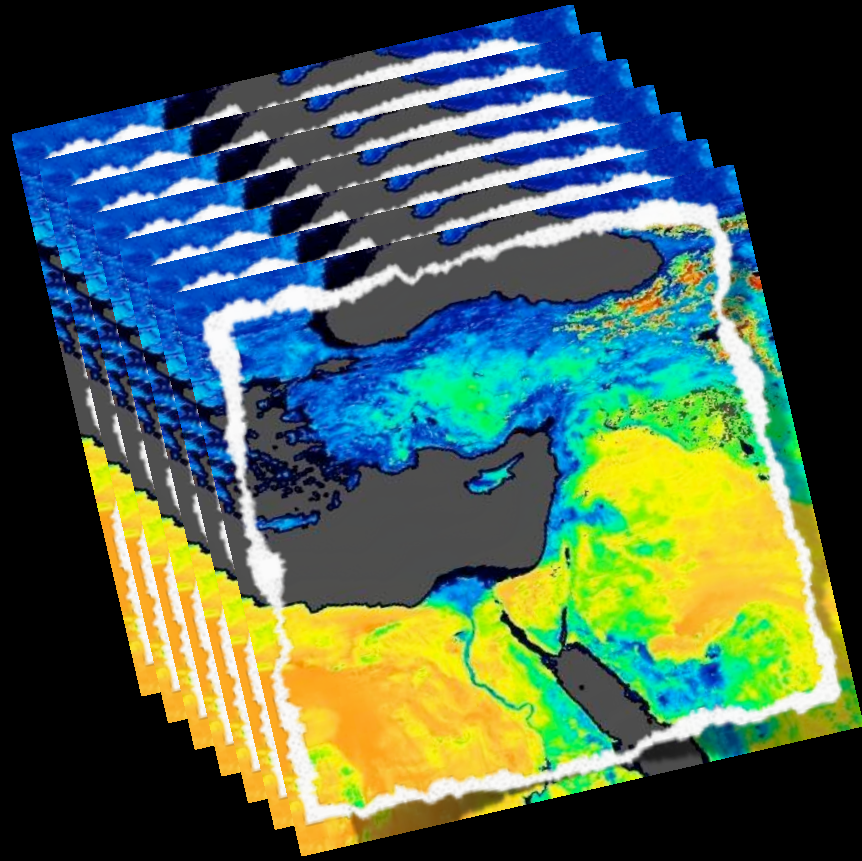
Self Organizing Map Classification



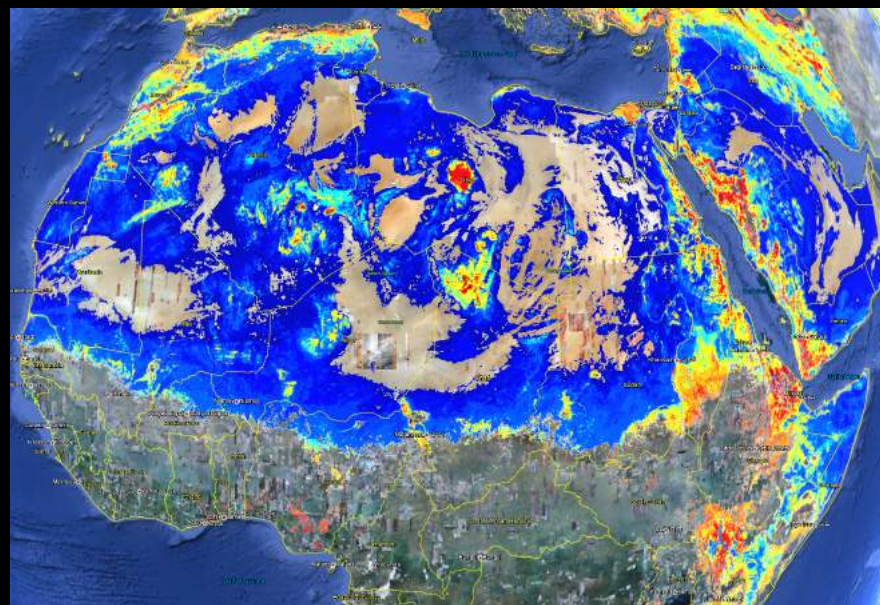
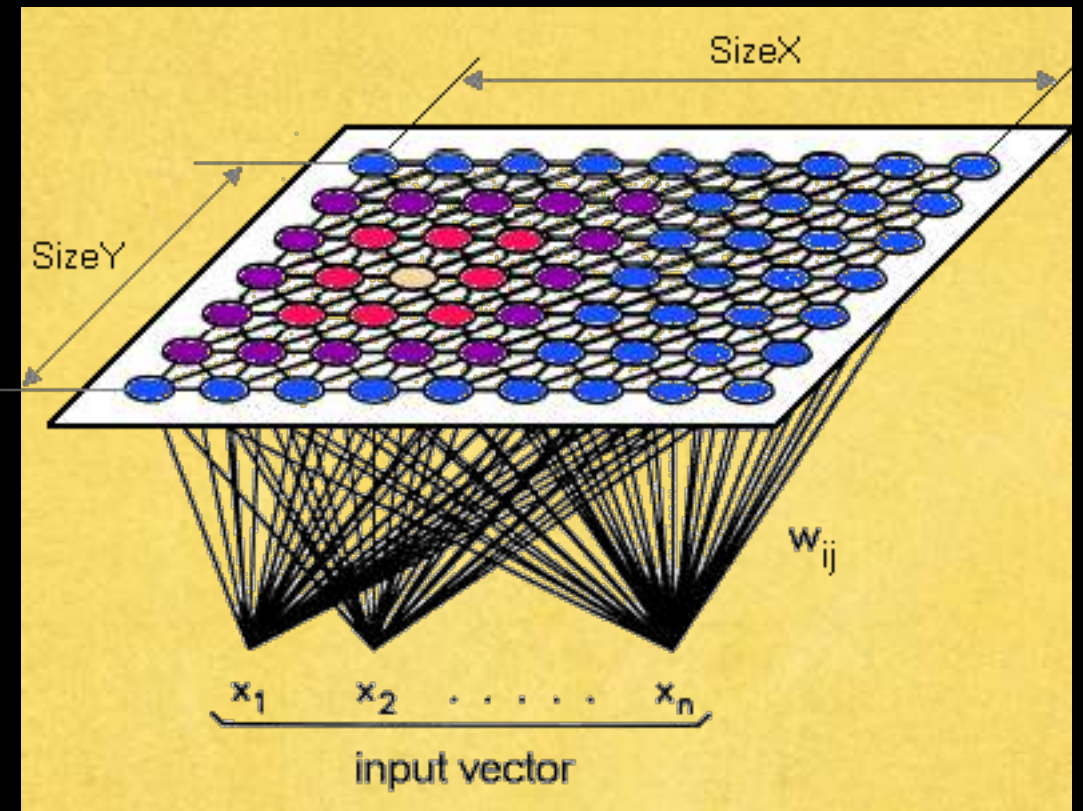
7 Bands
MODIS MCD43C3
bihemispherical reflectance



Self Organizing Map Classification



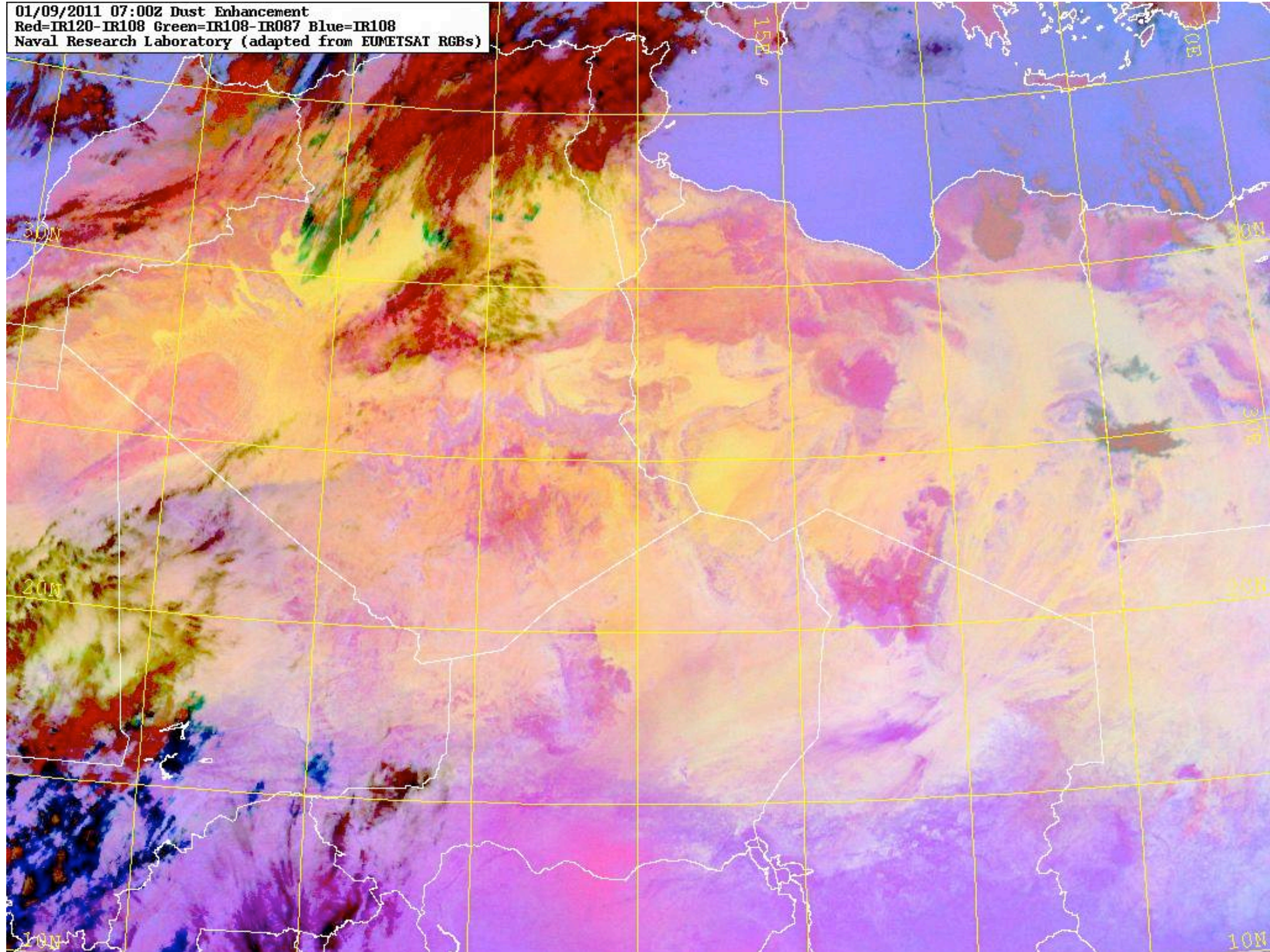
7 Bands
MODIS MCD43C3
bihemispherical reflectance



Chad: Bodélé Depression

Dust Event: March 16, 2010 (7Z -12Z)

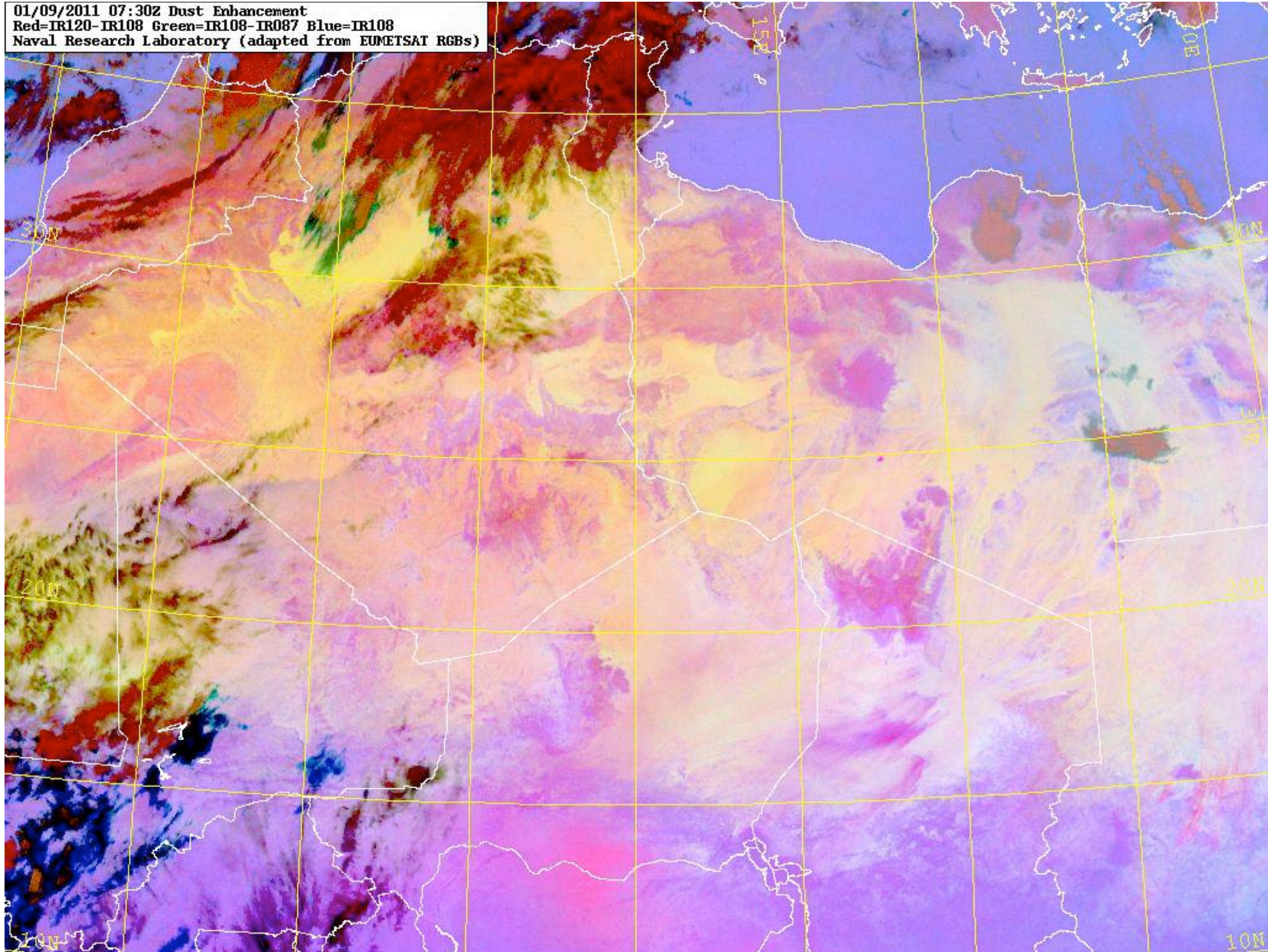
Located at the southern edge of the Sahara Desert in north central Africa, is the lowest point in Chad. Dust storms from the Bodélé Depression occur on average about 100 days per year. The Bodélé depression is a single spot in the Sahara that provides most of the mineral dust to the Amazon forest.



Chad: Bodélé Depression

Dust Event: March 16, 2010 (7Z -12Z)

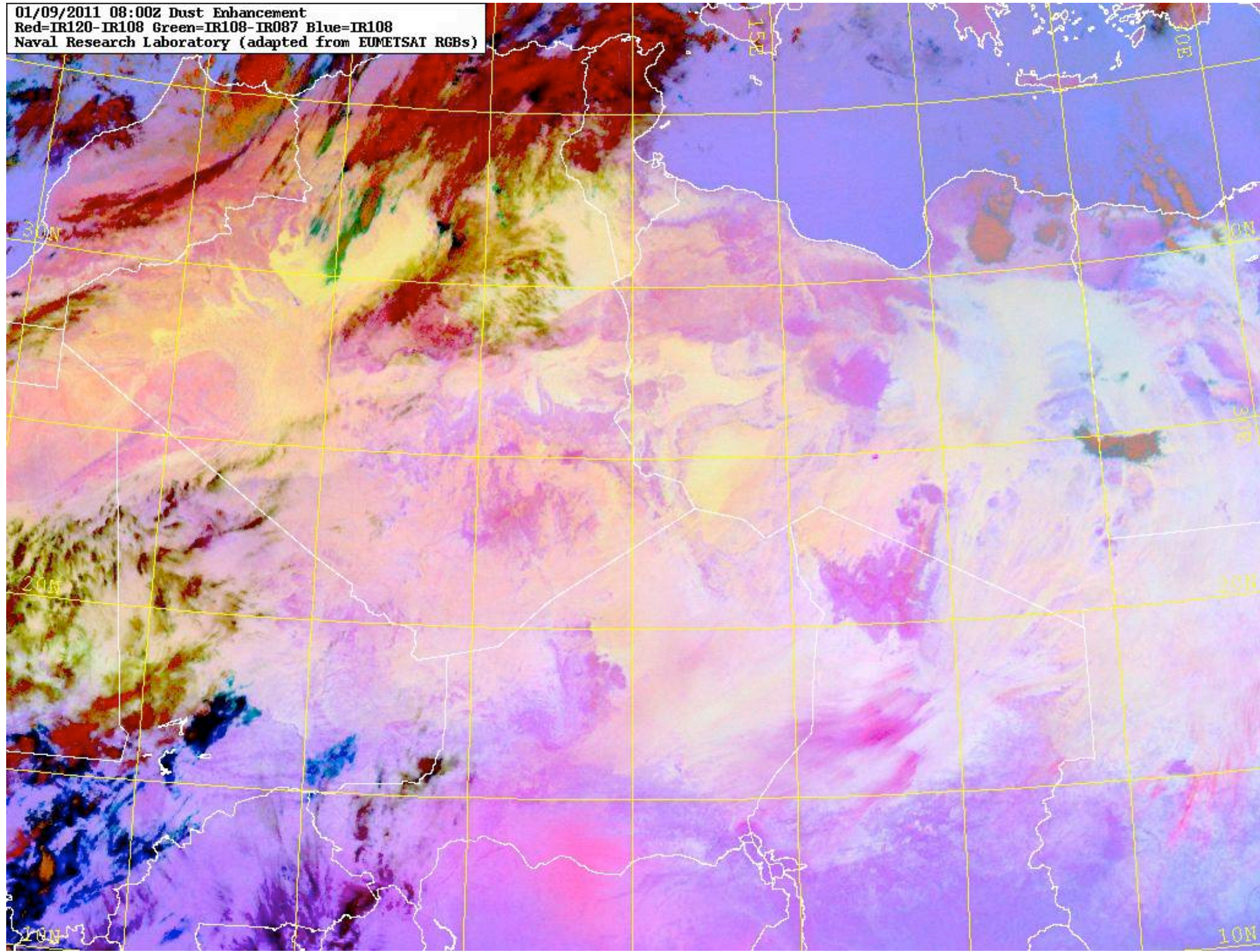
Located at the southern edge of the Sahara Desert in north central Africa, is the lowest point in Chad. Dust storms from the Bodélé Depression occur on average about 100 days per year. The Bodélé depression is a single spot in the Sahara that provides most of the mineral dust to the Amazon forest.



Chad: Bodélé Depression

Dust Event: March 16, 2010 (7Z -12Z)

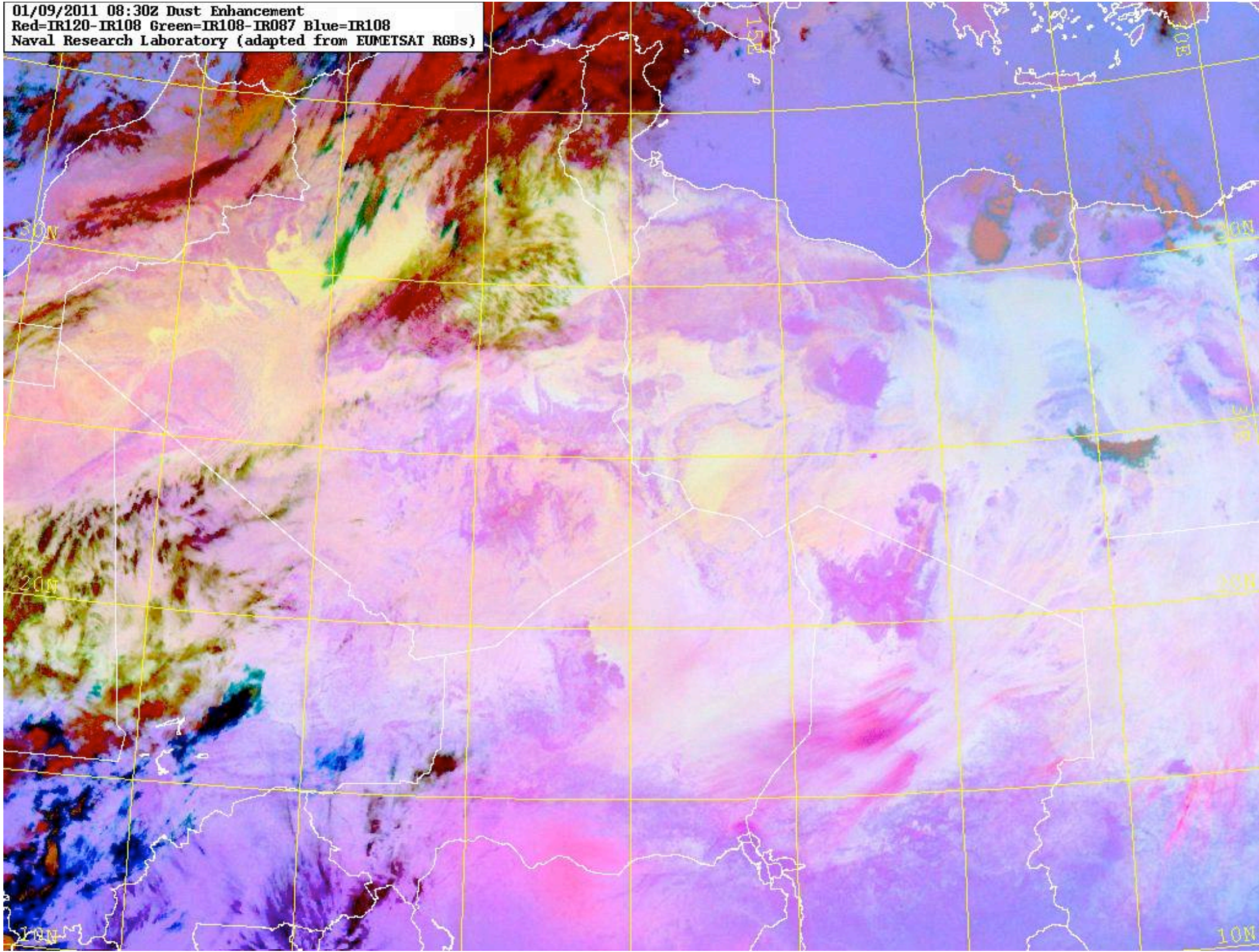
Located at the southern edge of the Sahara Desert in north central Africa, is the lowest point in Chad. Dust storms from the Bodélé Depression occur on average about 100 days per year. The Bodélé depression is a single spot in the Sahara that provides most of the mineral dust to the Amazon forest.



Chad: Bodélé Depression

Dust Event: March 16, 2010 (7Z -12Z)

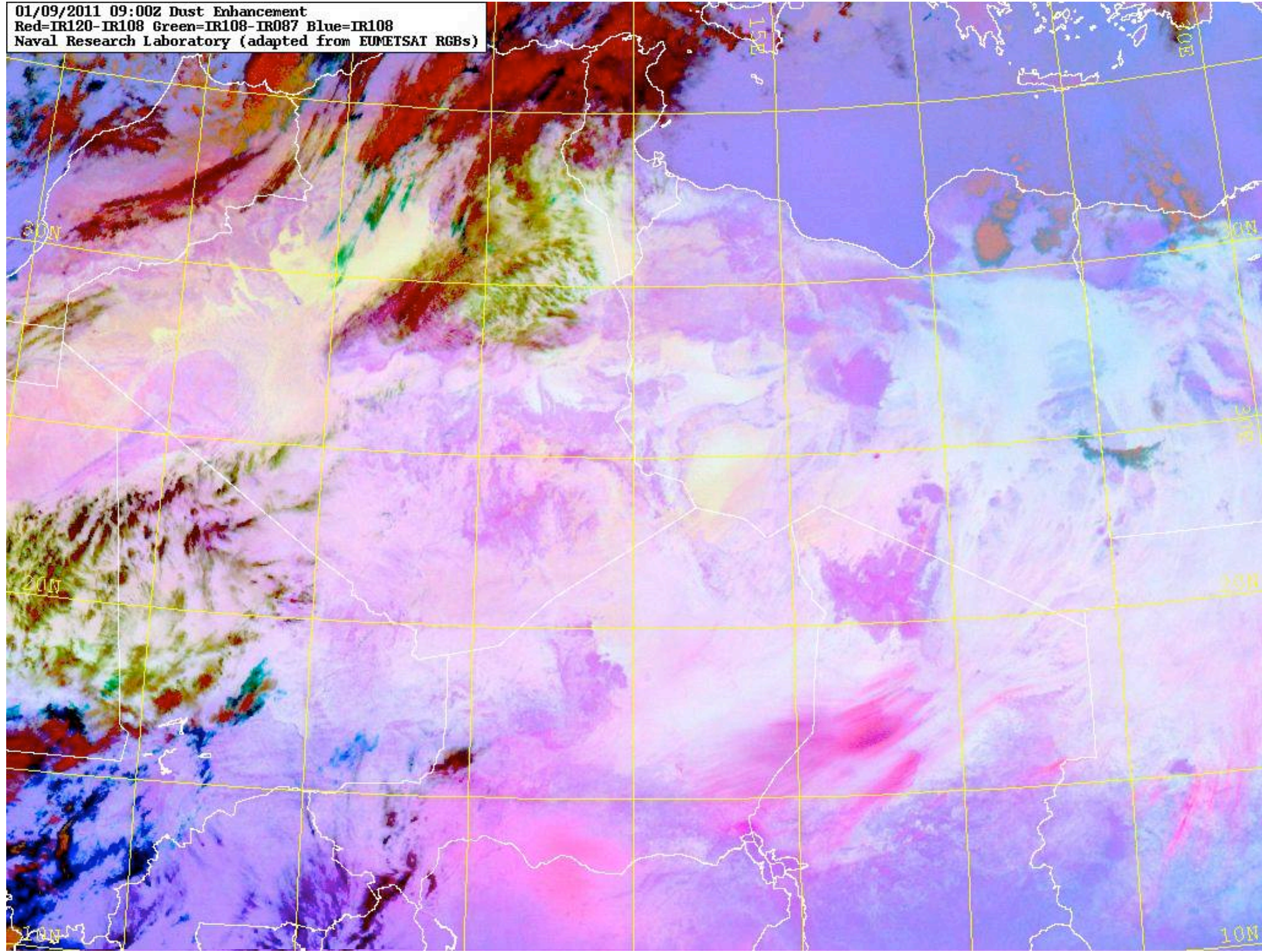
Located at the southern edge of the Sahara Desert in north central Africa, is the lowest point in Chad. Dust storms from the Bodélé Depression occur on average about 100 days per year. The Bodélé depression is a single spot in the Sahara that provides most of the mineral dust to the Amazon forest.



Chad: Bodélé Depression

Dust Event: March 16, 2010 (7Z -12Z)

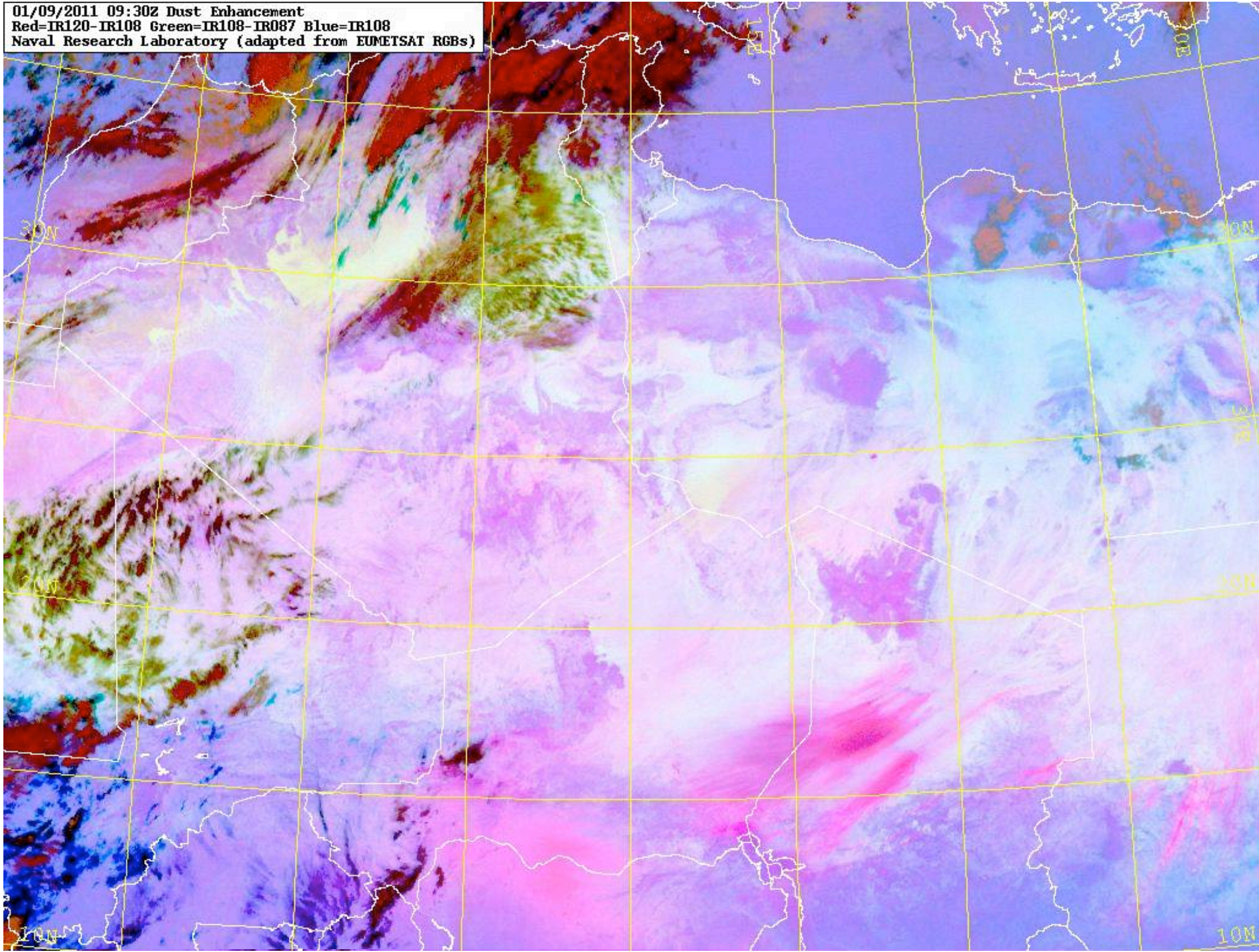
Located at the southern edge of the Sahara Desert in north central Africa, is the lowest point in Chad. Dust storms from the Bodélé Depression occur on average about 100 days per year. The Bodélé depression is a single spot in the Sahara that provides most of the mineral dust to the Amazon forest.



Chad: Bodélé Depression

Dust Event: March 16, 2010 (7Z -12Z)

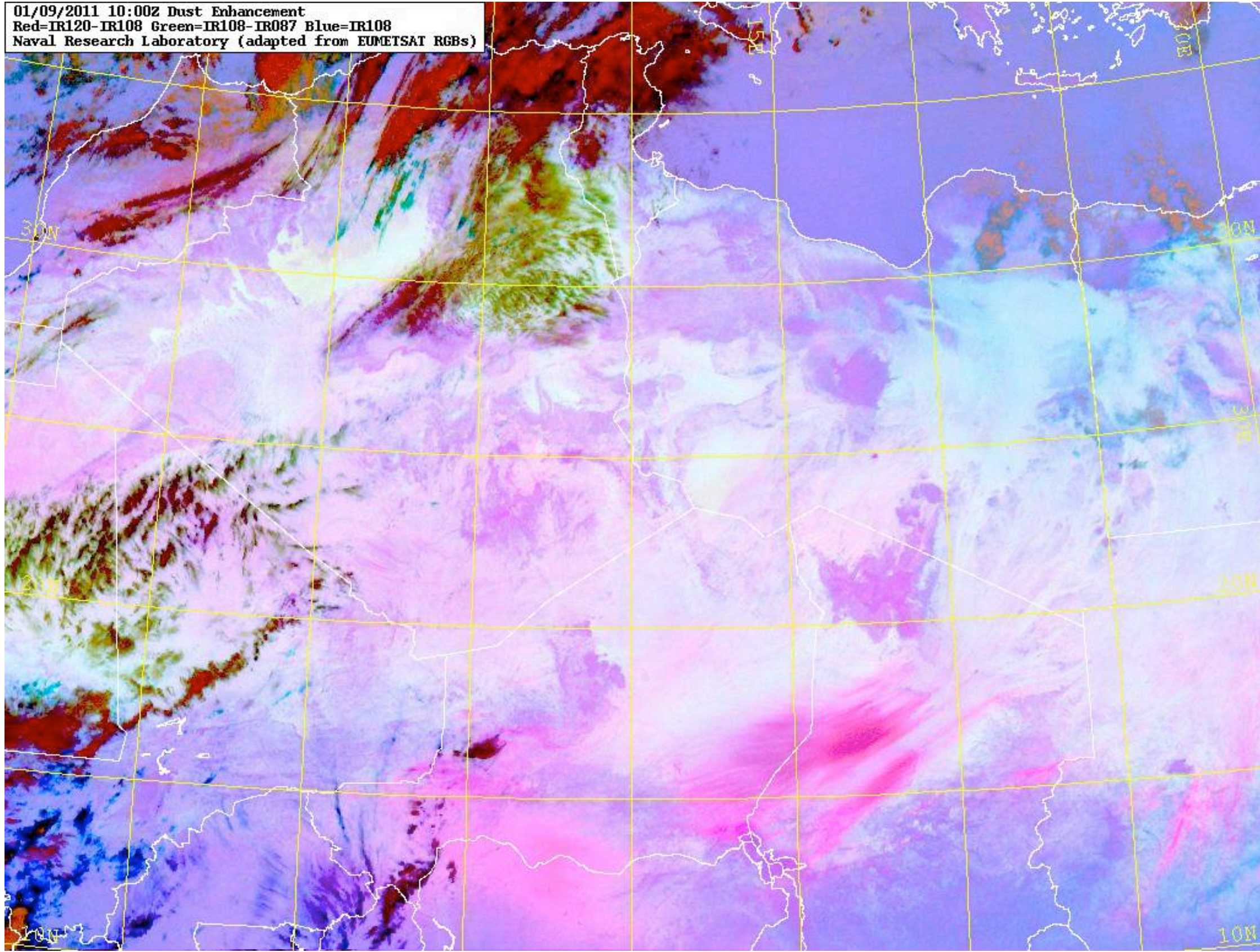
Located at the southern edge of the Sahara Desert in north central Africa, is the lowest point in Chad. Dust storms from the Bodélé Depression occur on average about 100 days per year. The Bodélé depression is a single spot in the Sahara that provides most of the mineral dust to the Amazon forest.



Chad: Bodélé Depression

Dust Event: March 16, 2010 (7Z -12Z)

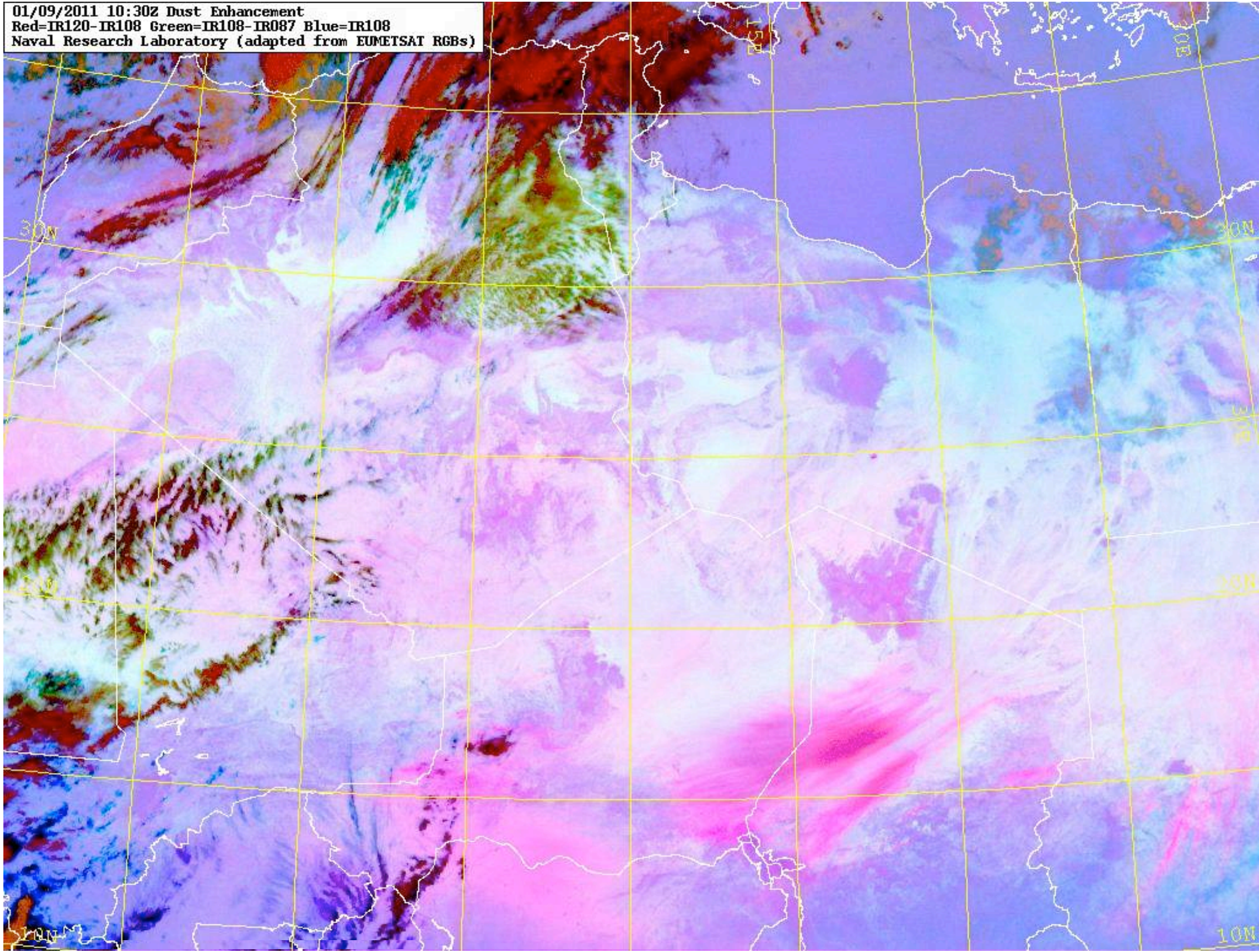
Located at the southern edge of the Sahara Desert in north central Africa, is the lowest point in Chad. Dust storms from the Bodélé Depression occur on average about 100 days per year. The Bodélé depression is a single spot in the Sahara that provides most of the mineral dust to the Amazon forest.



Chad: Bodélé Depression

Dust Event: March 16, 2010 (7Z -12Z)

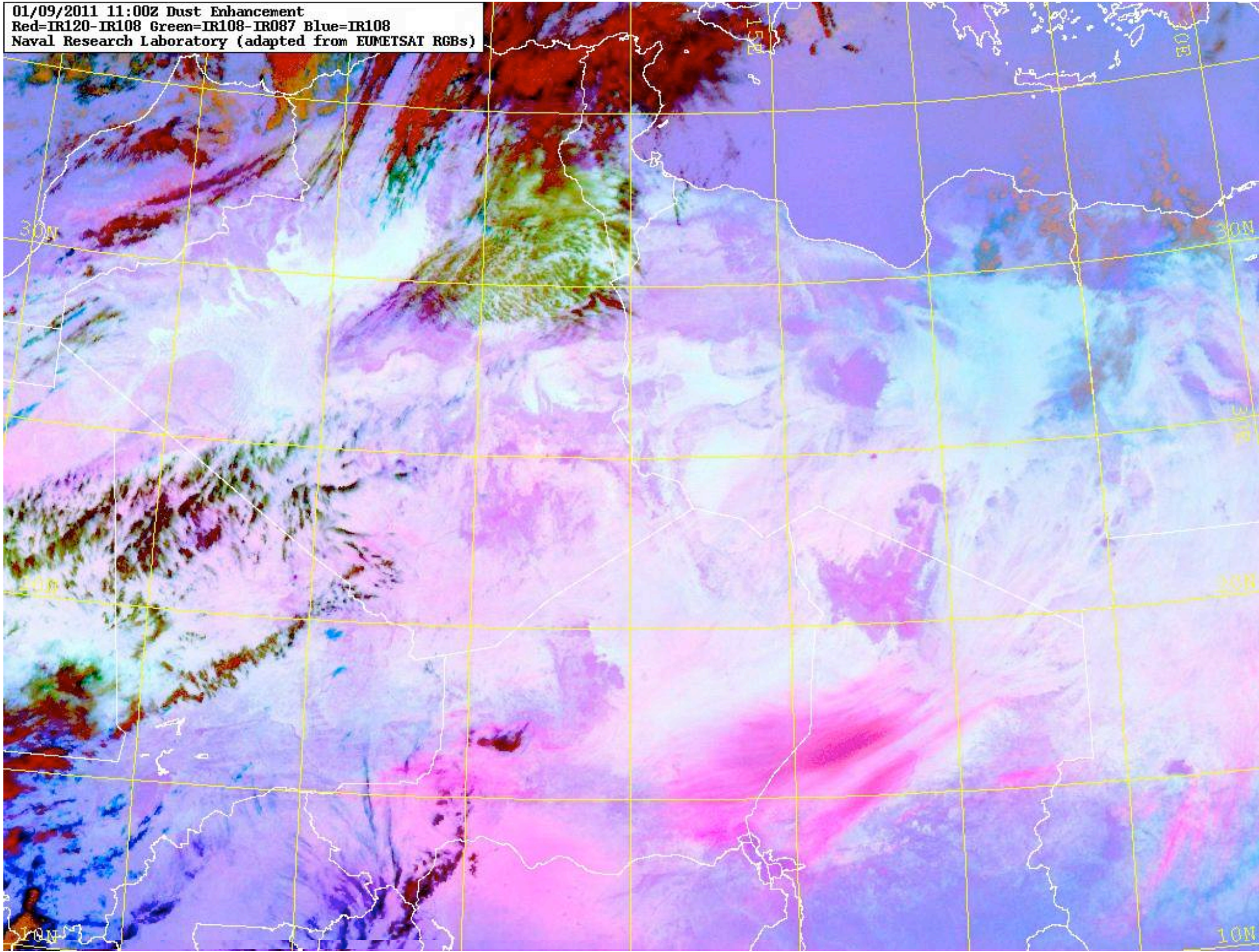
Located at the southern edge of the Sahara Desert in north central Africa, is the lowest point in Chad. Dust storms from the Bodélé Depression occur on average about 100 days per year. The Bodélé depression is a single spot in the Sahara that provides most of the mineral dust to the Amazon forest.



Chad: Bodélé Depression

Dust Event: March 16, 2010 (7Z -12Z)

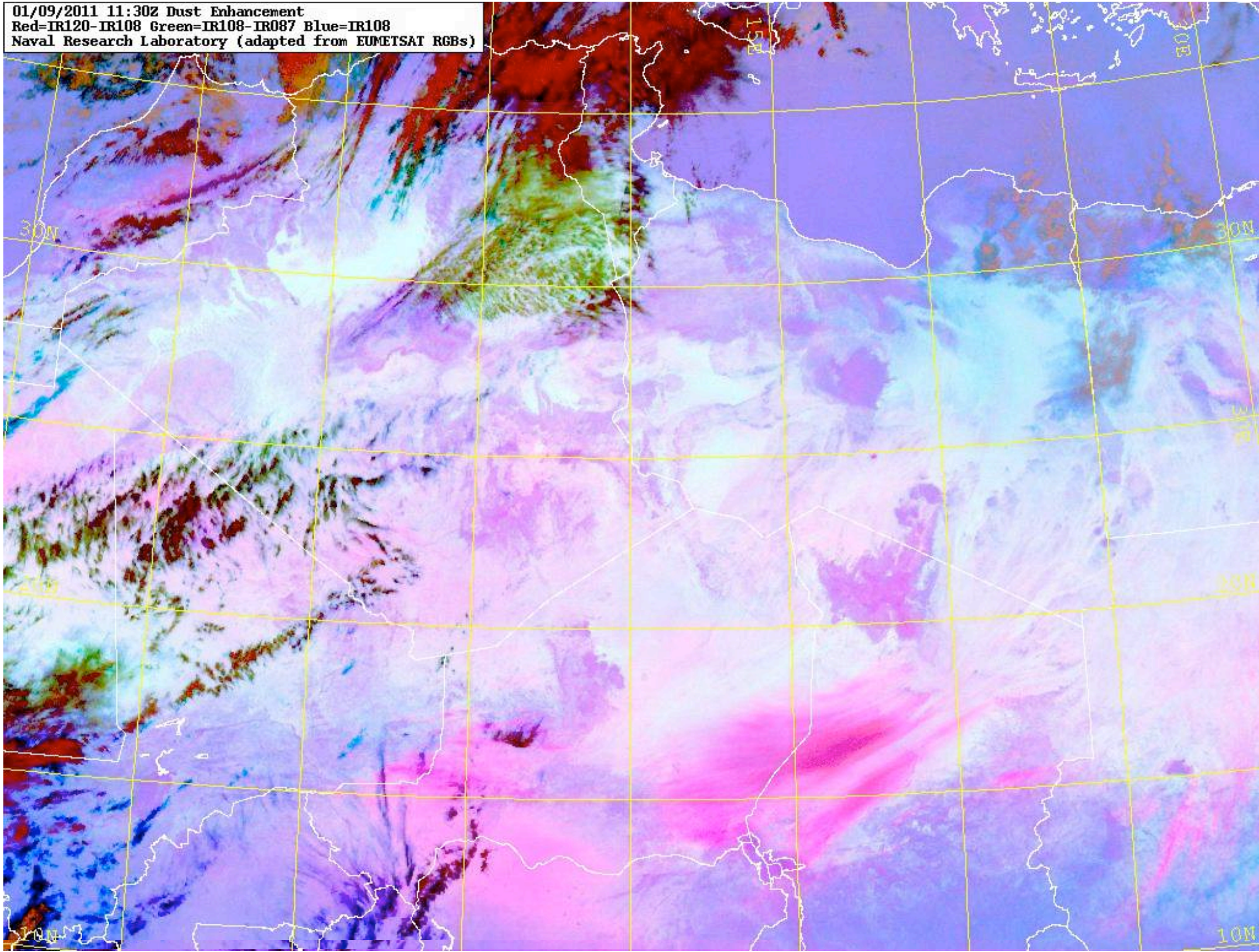
Located at the southern edge of the Sahara Desert in north central Africa, is the lowest point in Chad. Dust storms from the Bodélé Depression occur on average about 100 days per year. The Bodélé depression is a single spot in the Sahara that provides most of the mineral dust to the Amazon forest.



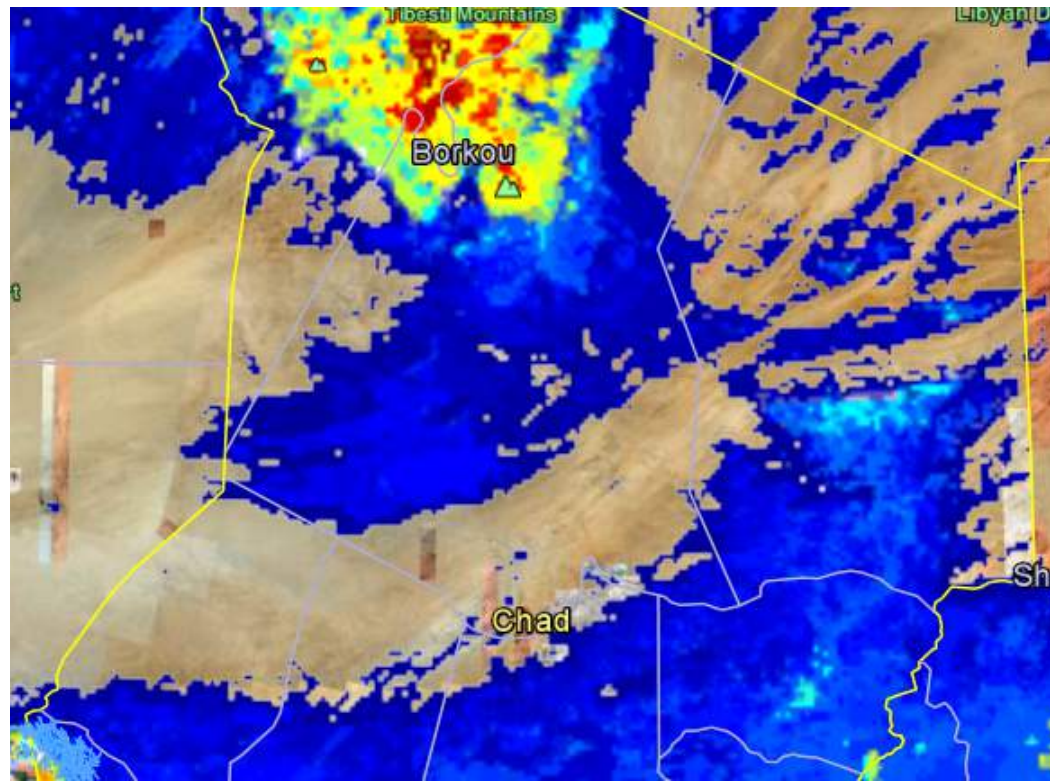
Chad: Bodélé Depression

Dust Event: March 16, 2010 (7Z -12Z)

Located at the southern edge of the Sahara Desert in north central Africa, is the lowest point in Chad. Dust storms from the Bodélé Depression occur on average about 100 days per year. The Bodélé depression is a single spot in the Sahara that provides most of the mineral dust to the Amazon forest.

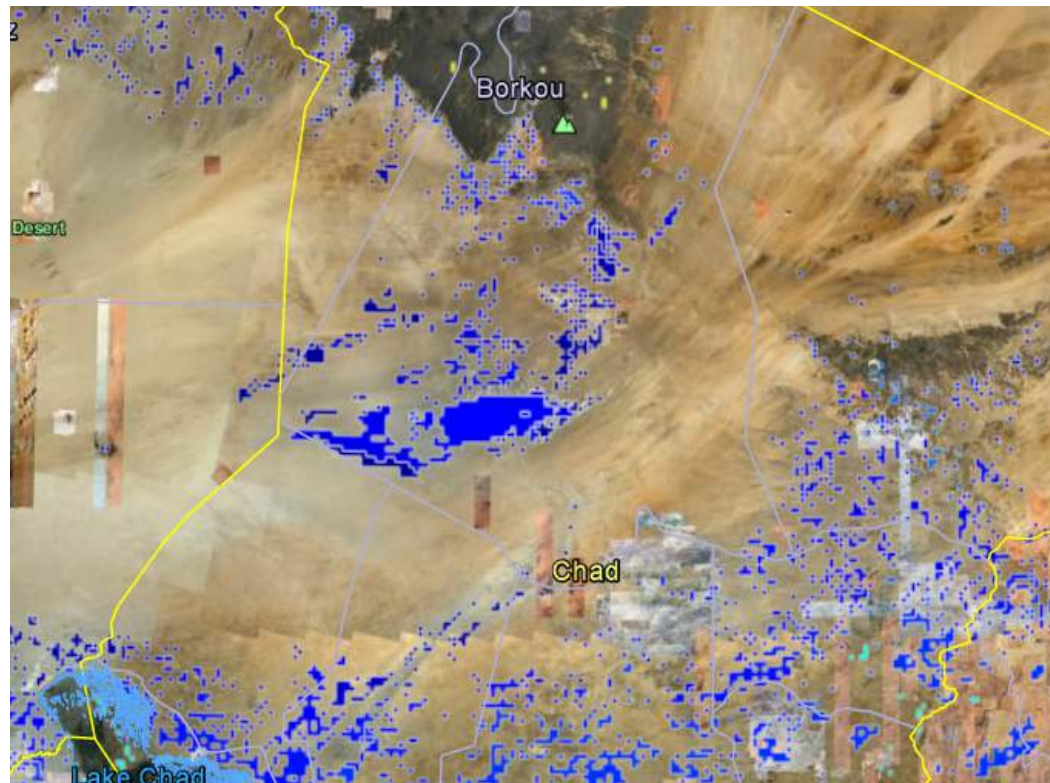


Chad: Bodélé Depression

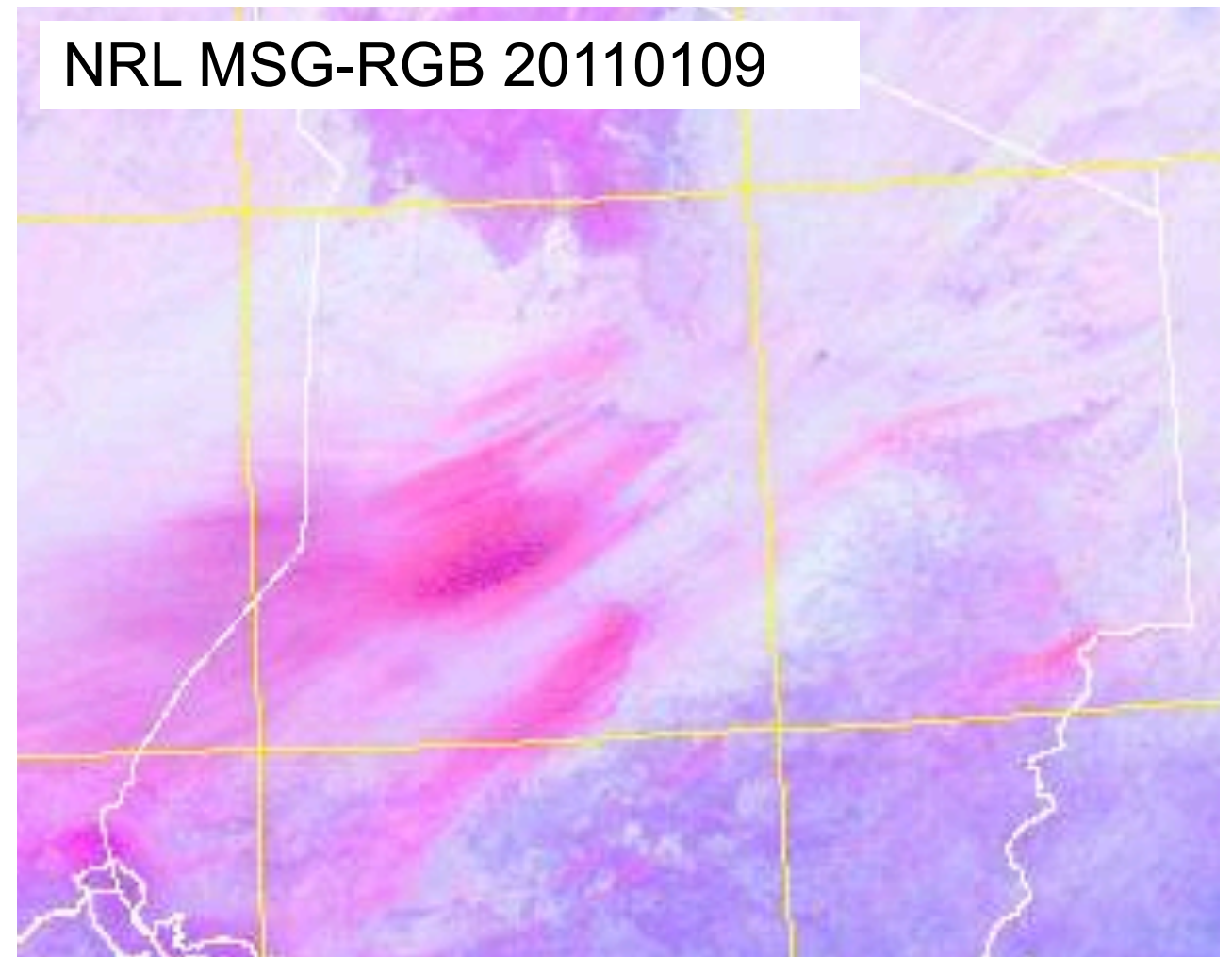


1000 SOM Classes

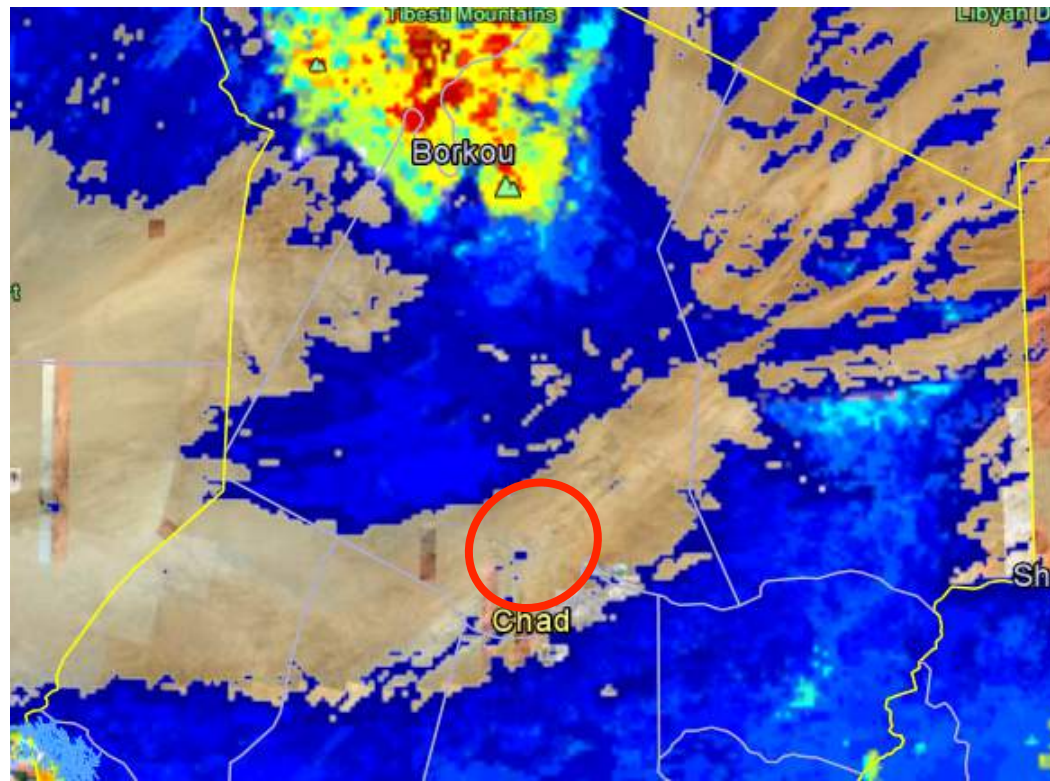
Source area is not designated in first pass of MODIS reflectance and land surface classification.



Selected SOM Classes

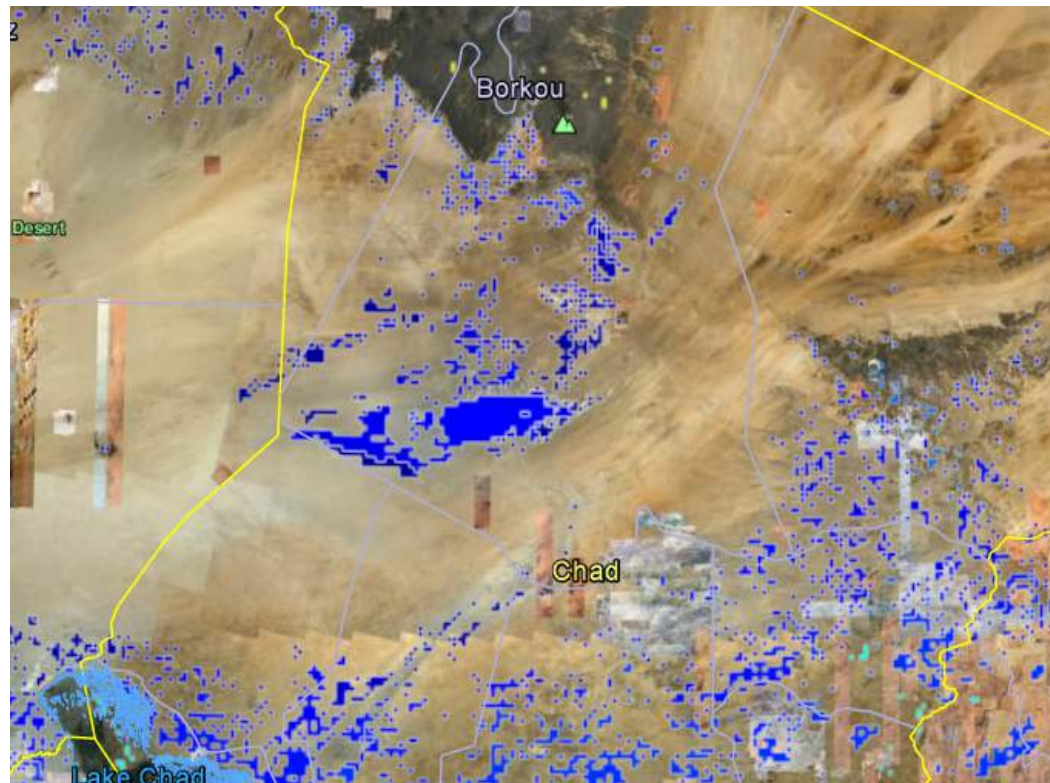


Chad: Bodélé Depression

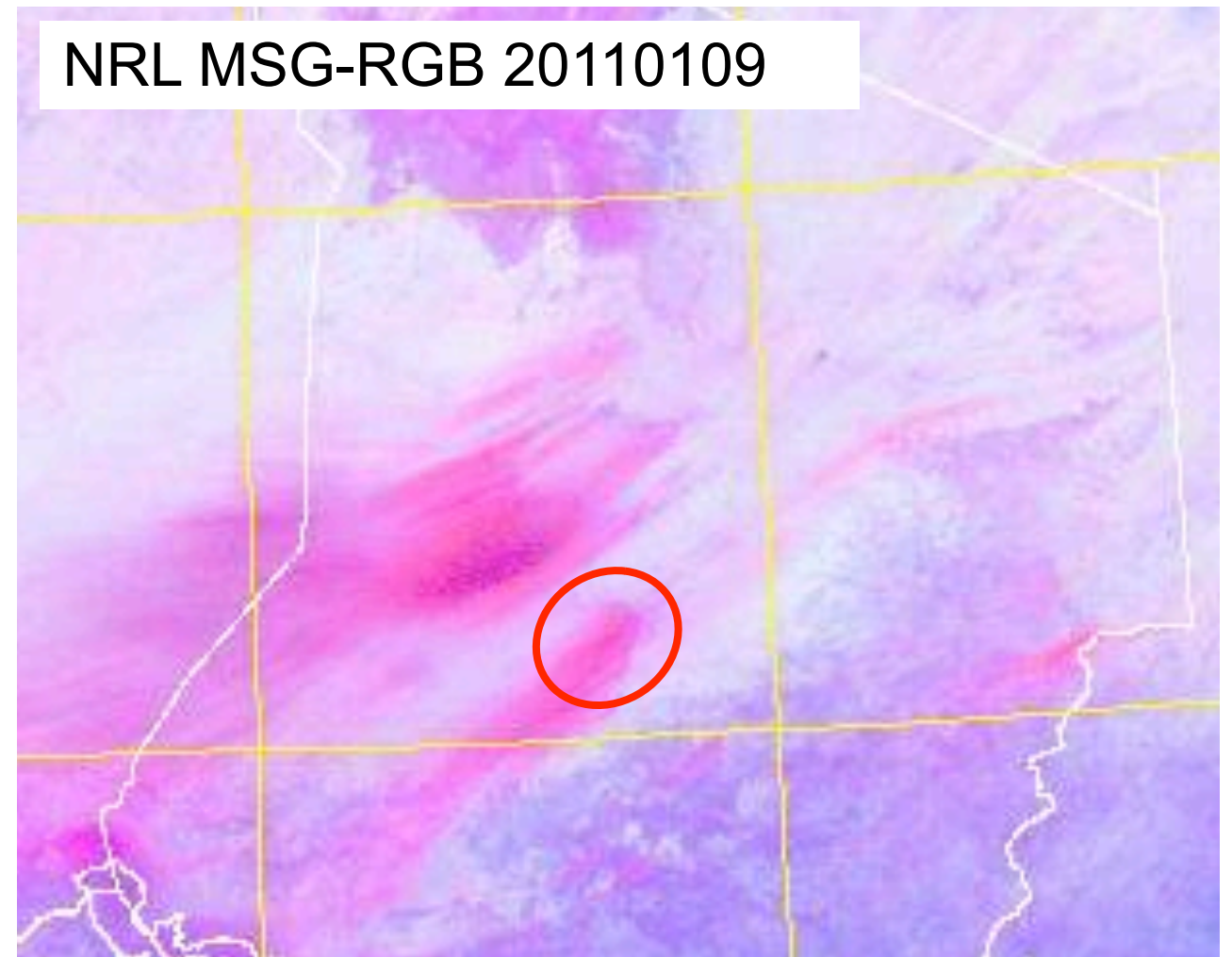


1000 SOM Classes

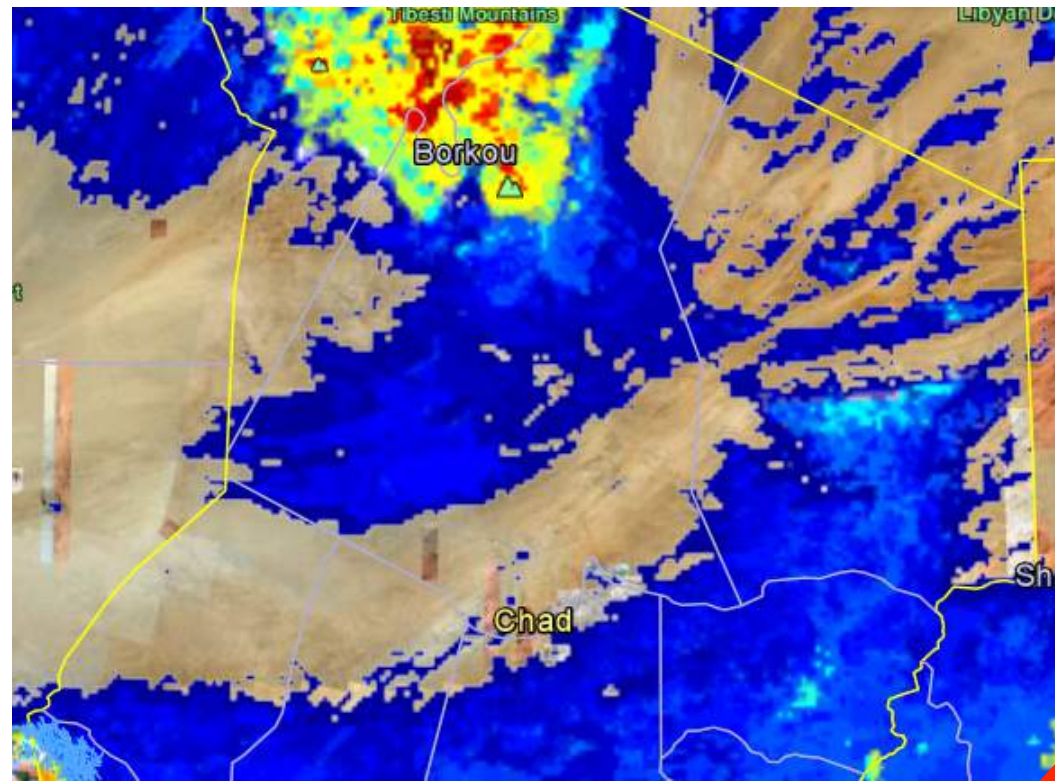
Source area is not designated in first pass of MODIS reflectance and land surface classification.



Selected SOM Classes

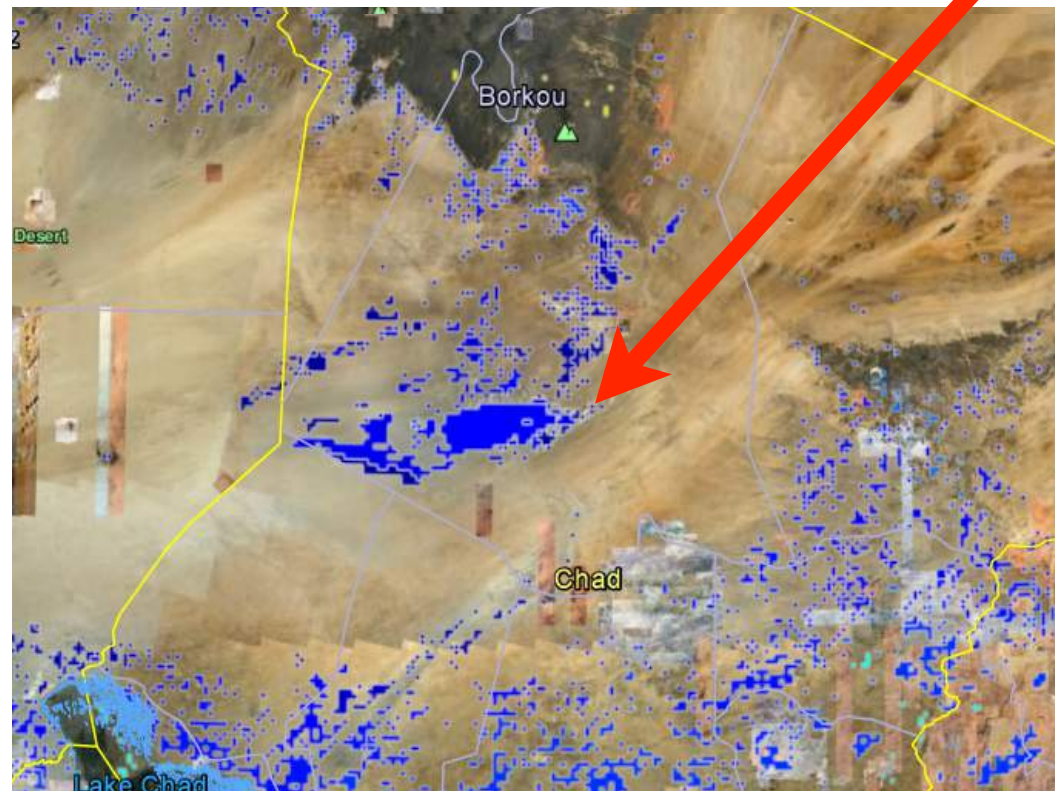
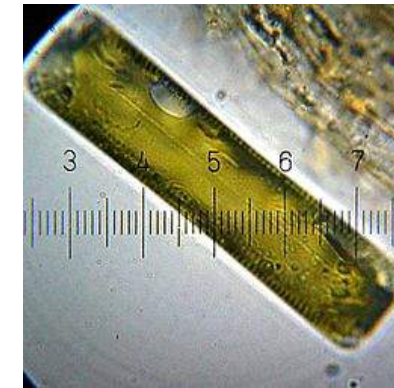
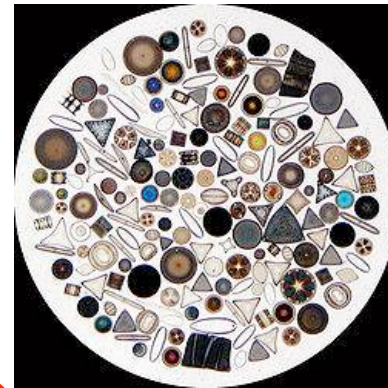


Chad: Bodélé Depression

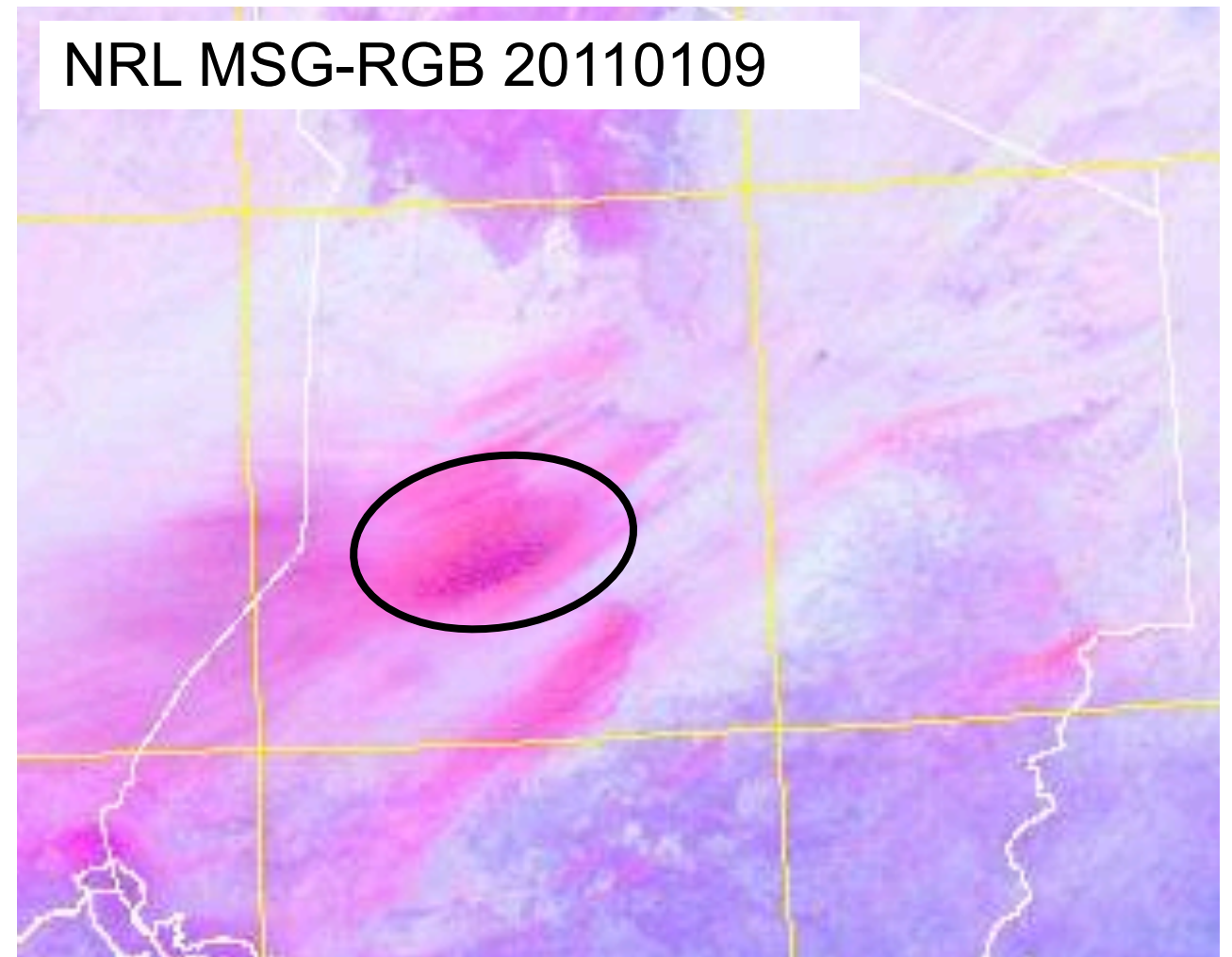


1000 SOM Classes

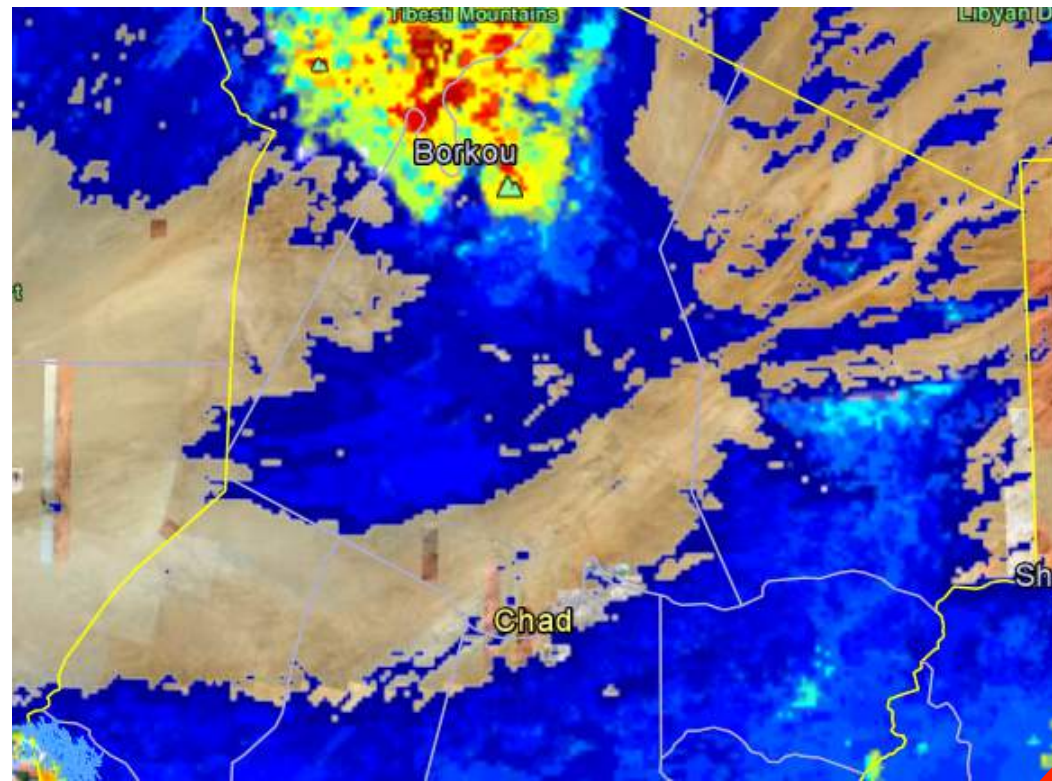
Class 137 maps diatom sediment in depression.



Selected Classes with Class 137

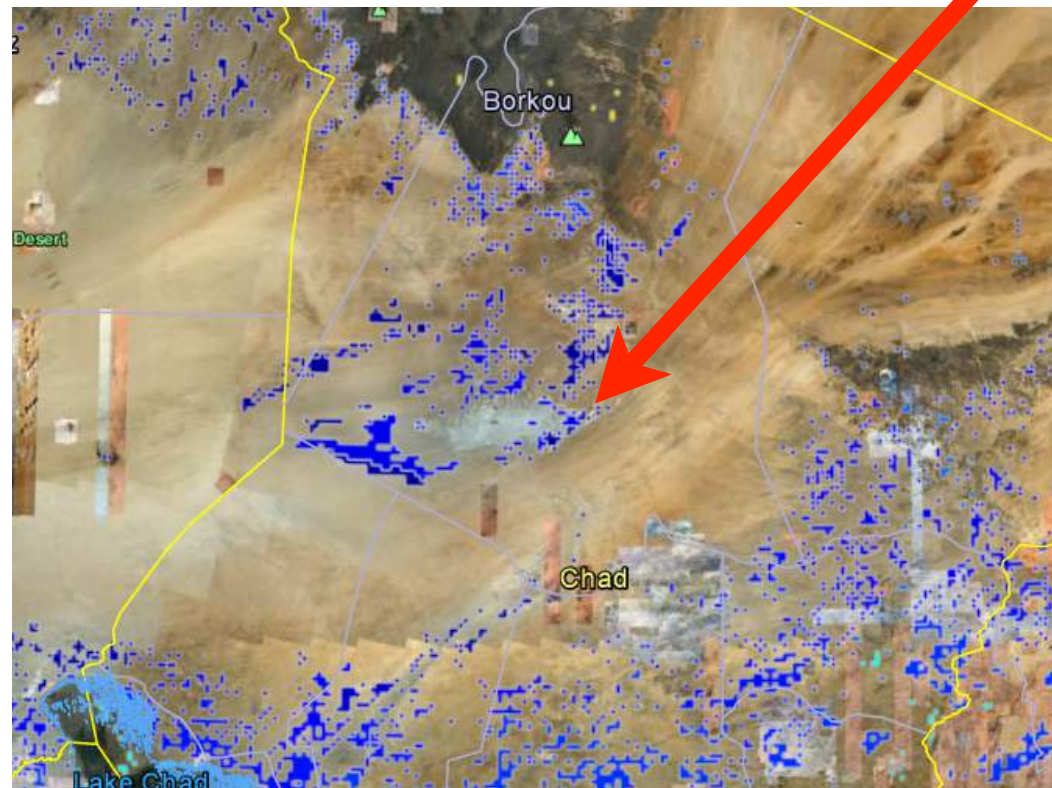
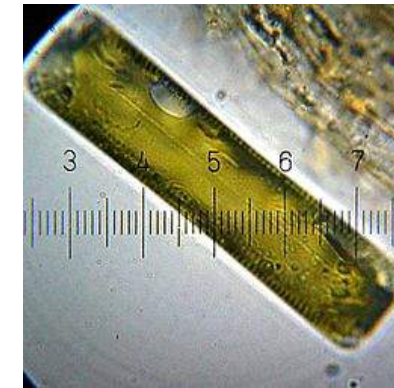
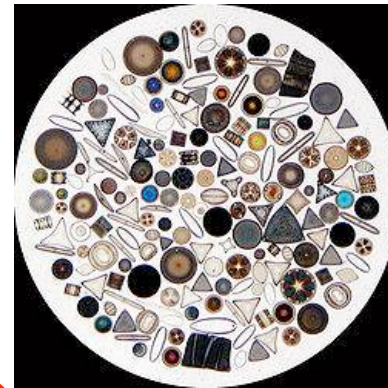


Chad: Bodélé Depression

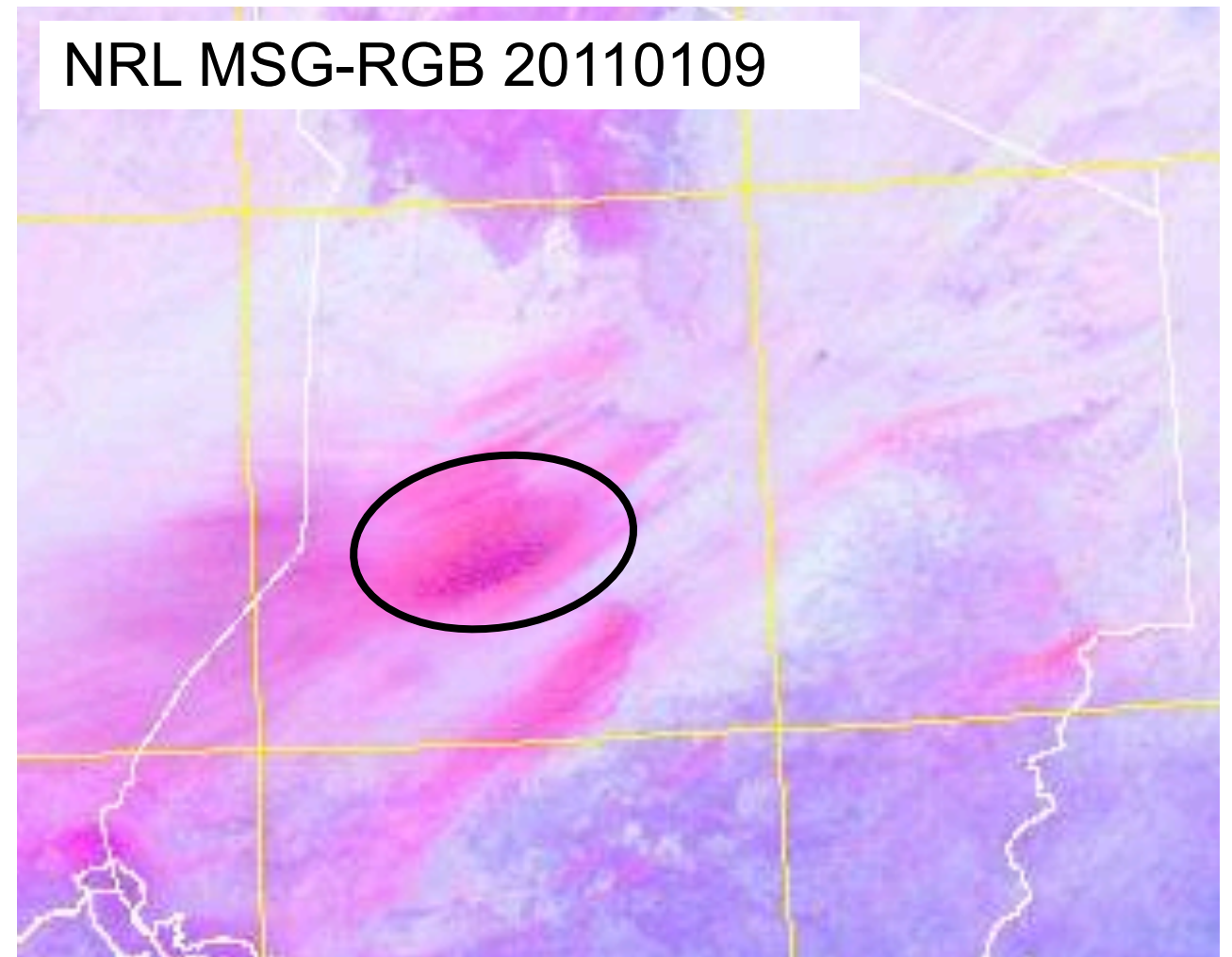


1000 SOM Classes

Class 137 maps diatom sediment in depression.

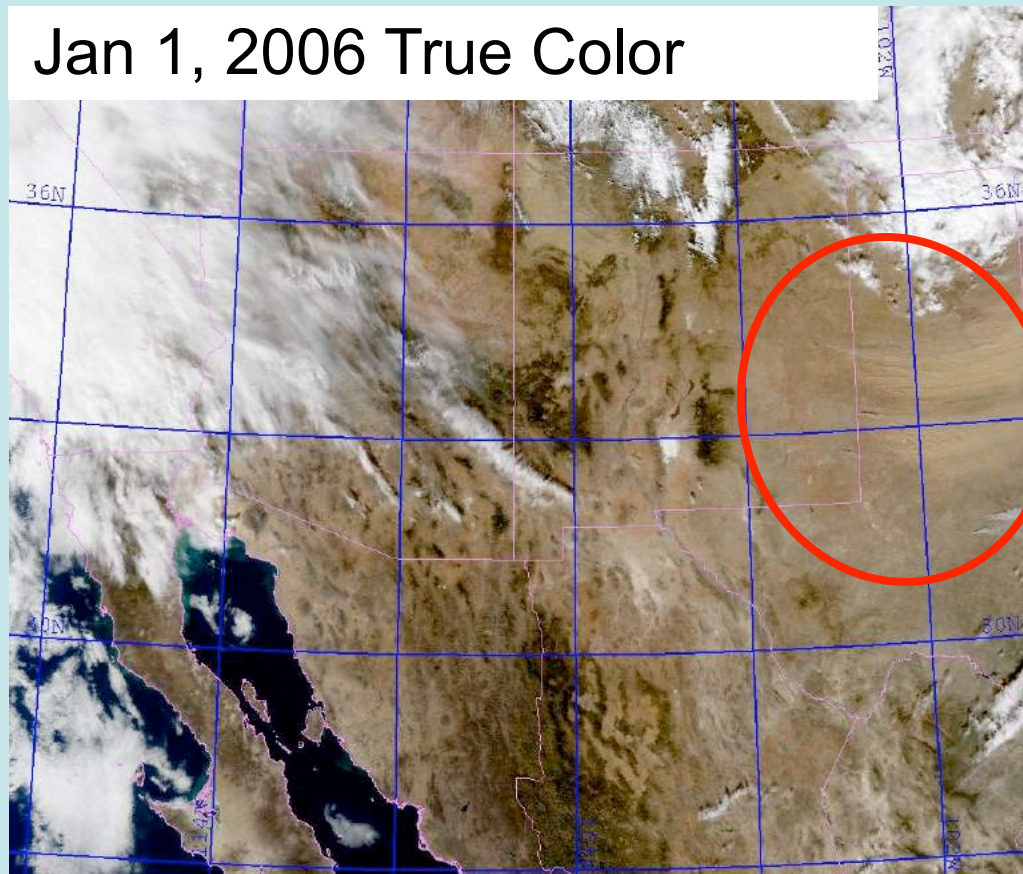


Selected Classes **Without Class 137**

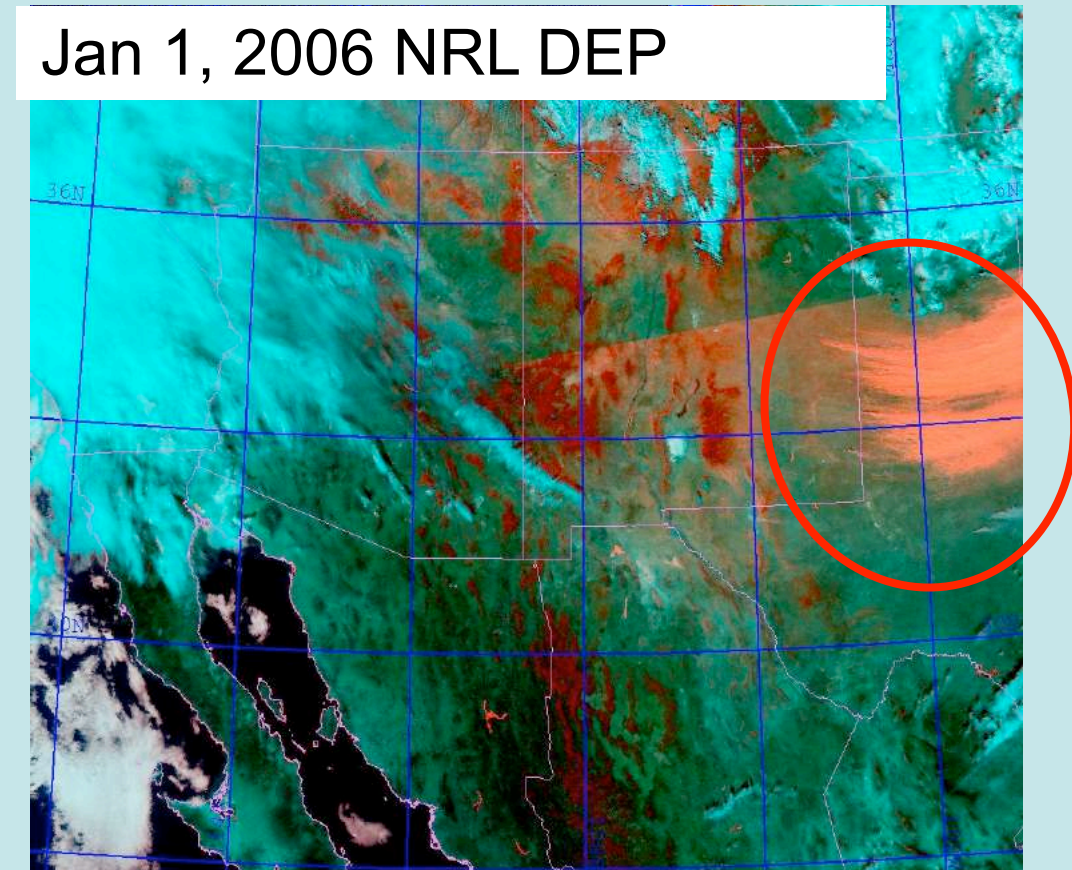


Sources along New Mexico/Texas border

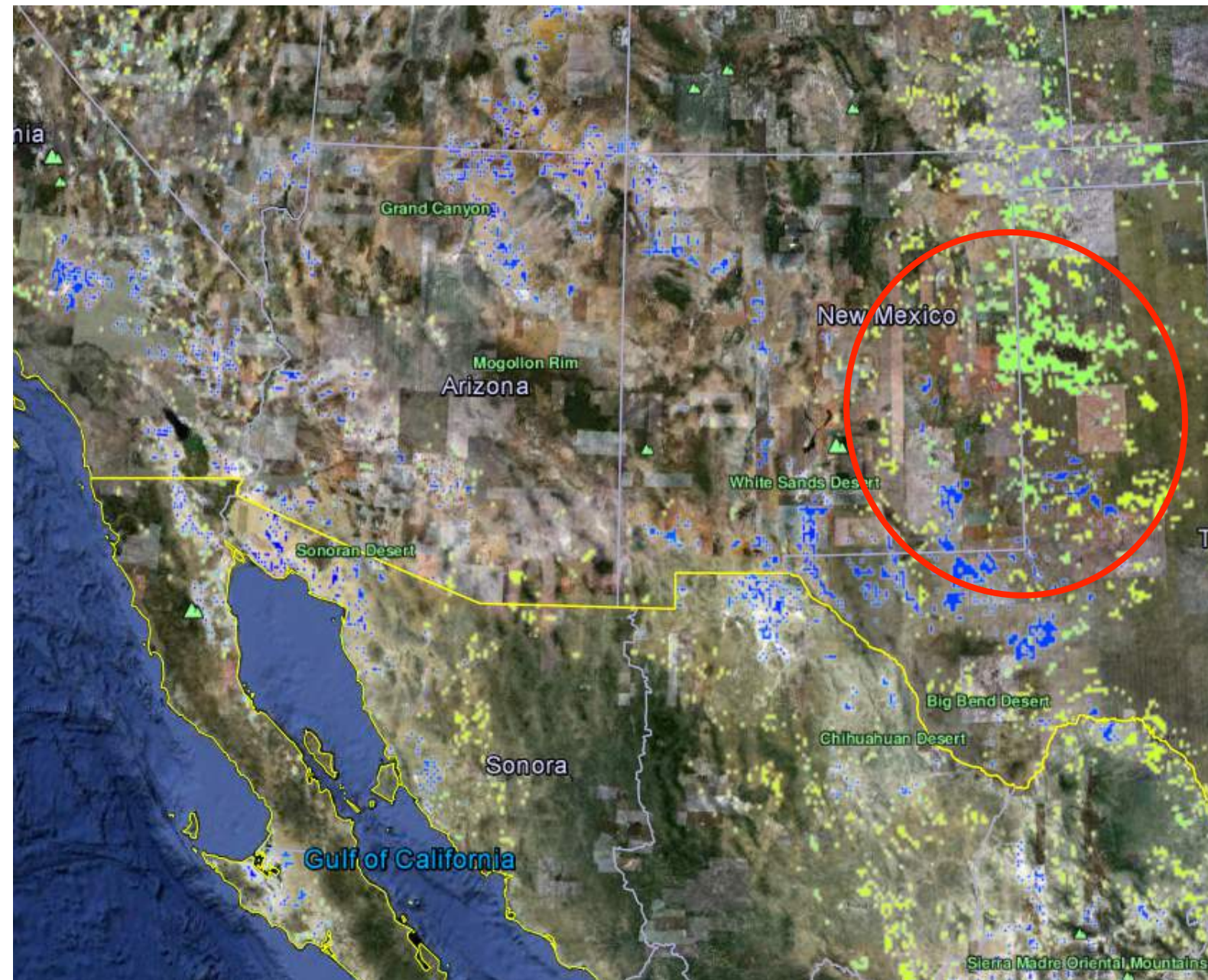
Jan 1, 2006 True Color



Jan 1, 2006 NRL DEP



Agricultural on high planes
Blue desert areas

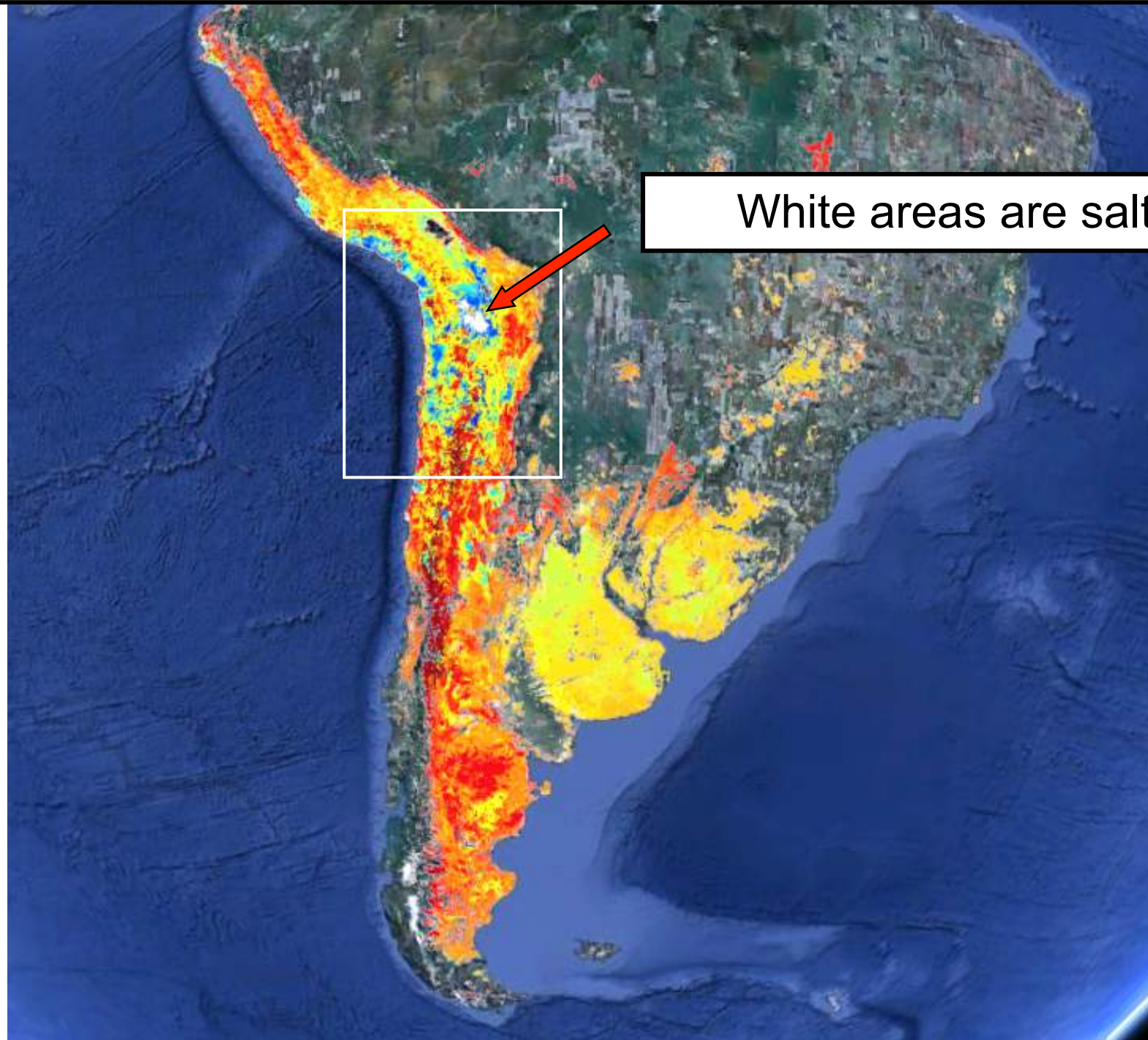


The North American sources have a different spectral signature than those we saw in SW Asia

All 1000-Classes mapped for South America



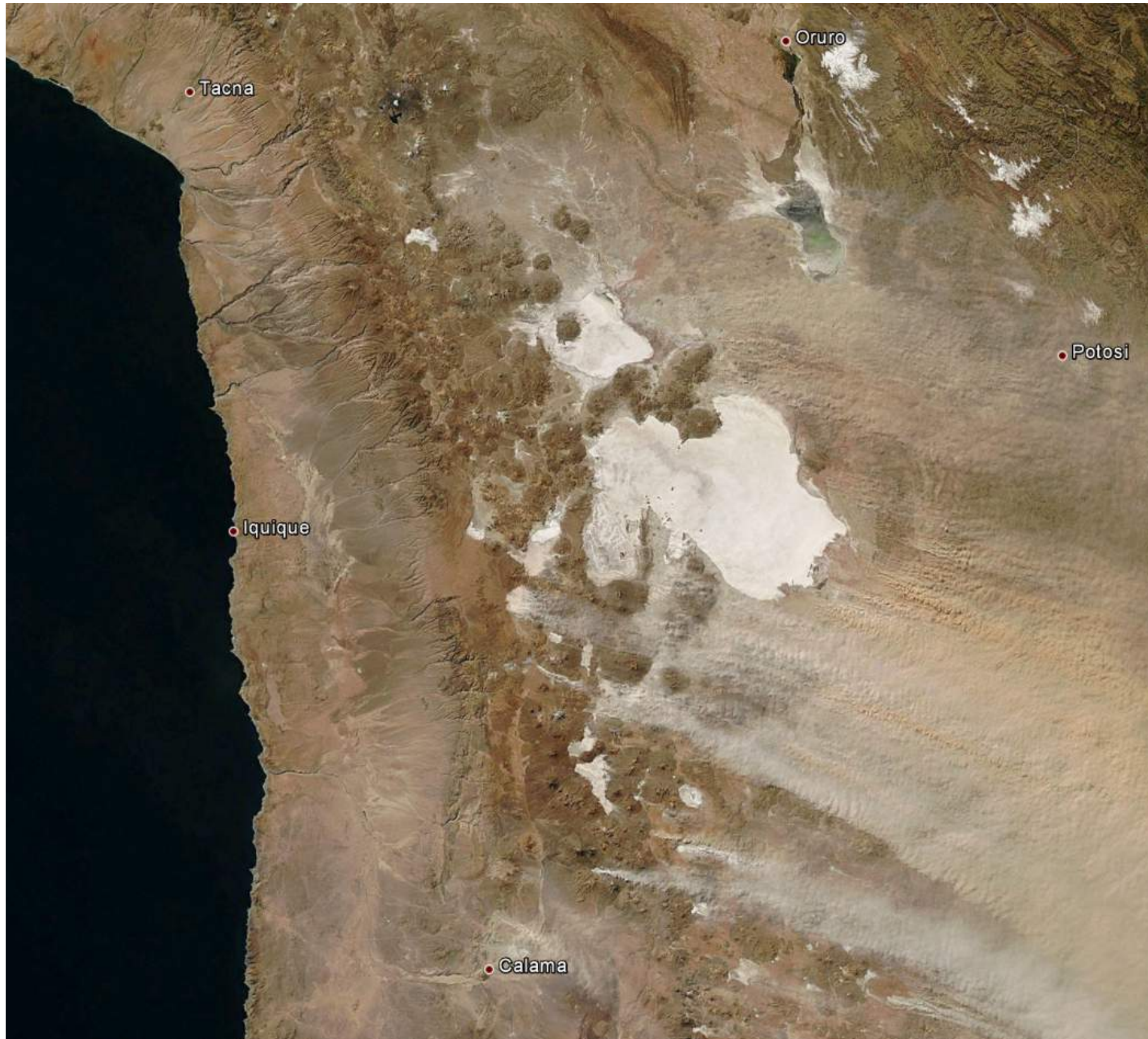
Blue colored SOM-Classes are concentrated in Atacama and Salar de Uyuni deserts



White areas are salt flats

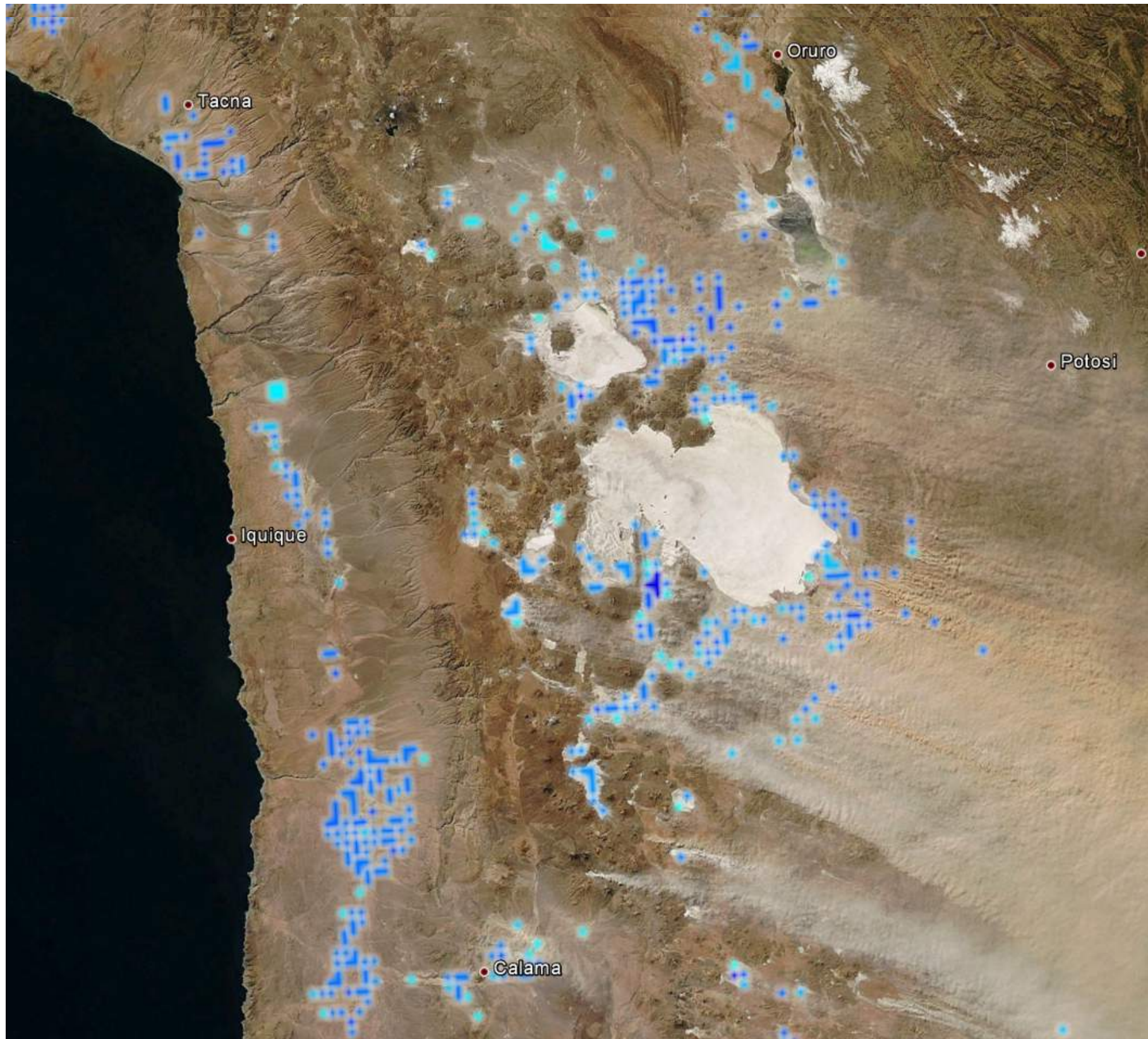
South America: Bolivia and Chile

July 18, 2010 MODIS Terra True Color



South America: Bolivia and Chile

Selected SOM-Classess in 200s, 300s, and 400s



GEOLOCATED ALLERGEN SENSING PLATFORM (GASP)

NSF FUNDING IS PROVIDING A CITY WIDE QUANTITATIVE VALIDATION CAMPAIGN OF OUR SMART CITY NETWORK OF IOT LASER BASED MINIATURE POLLEN SENSORS AND FACILITATE A LARGER SCALE ROLL OUT IN THE 200 US IGNITE CITIES ACROSS AMERICA.

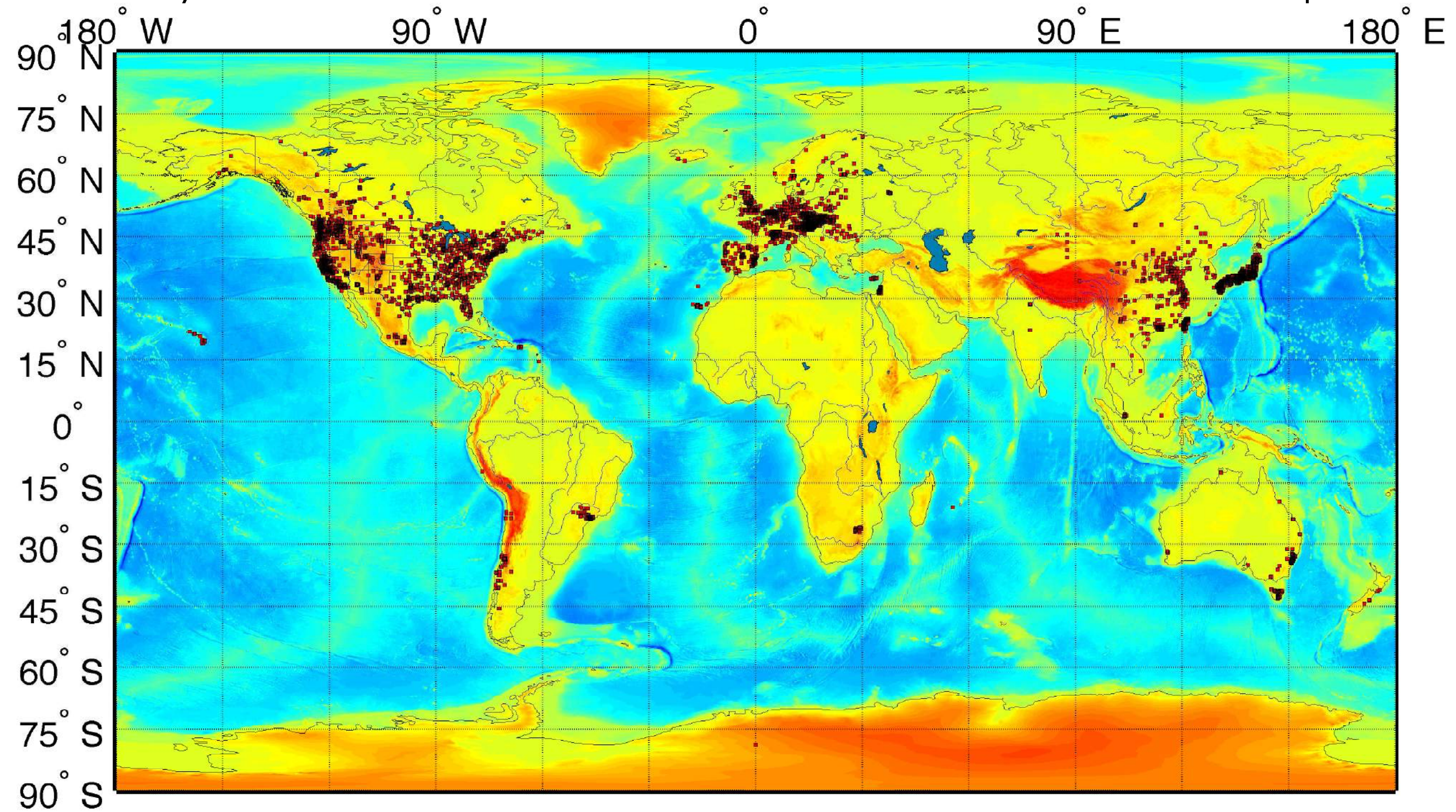
Why we care so much?

Approximately 50 million Americans have allergic diseases, including asthma and allergic rhinitis, both of which can be exacerbated by PM_{2.5}.

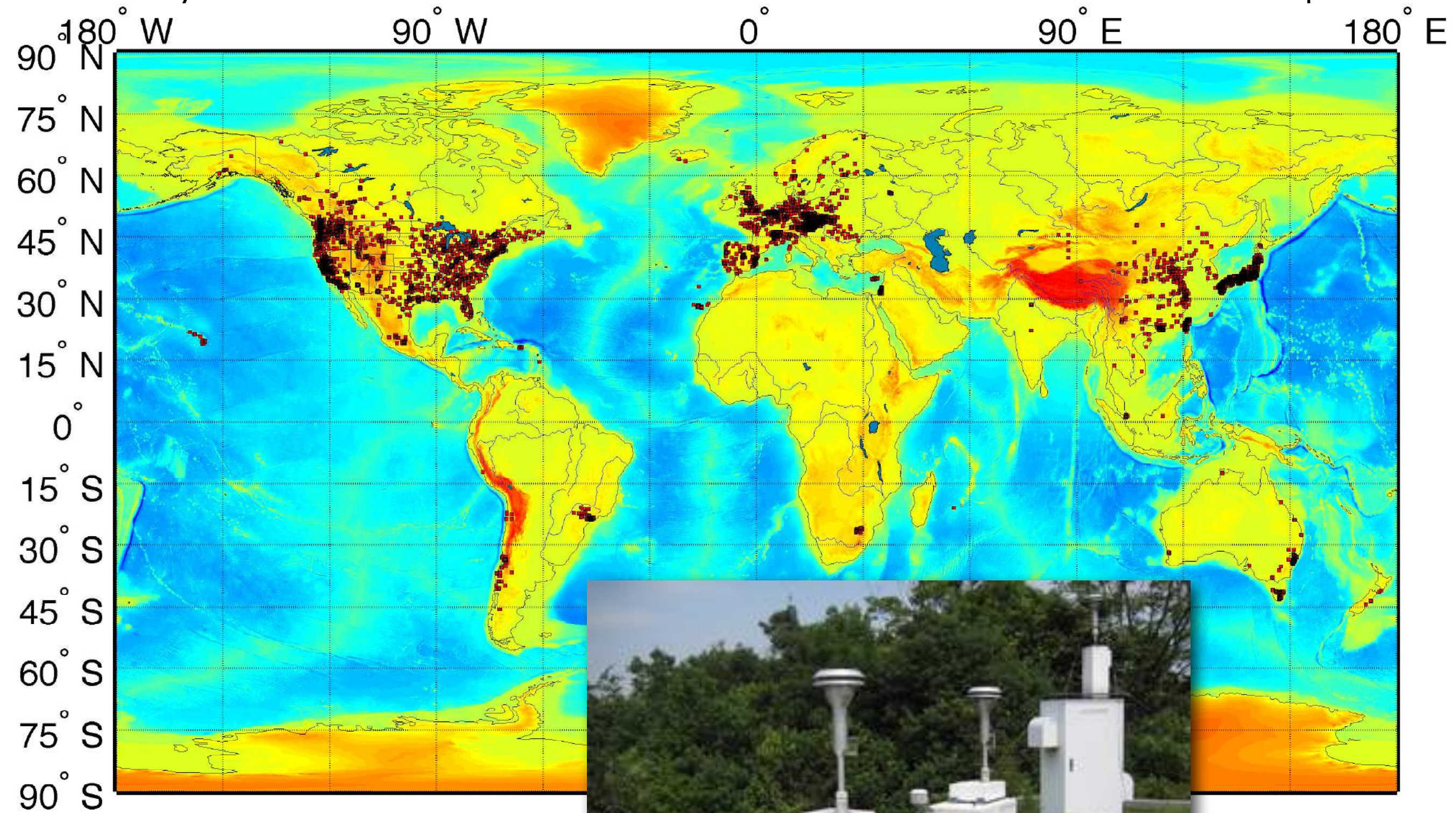
Every day in America 44,000 people have an asthma attack, and because of asthma 36,000 kids miss school, 27,000 adults miss work, 4,700 people visit the emergency room, 1,200 people are admitted to the hospital, and 9 people die.



Hourly Measurements from 55 countries and more than 8,000 measurement sites from 1997-present

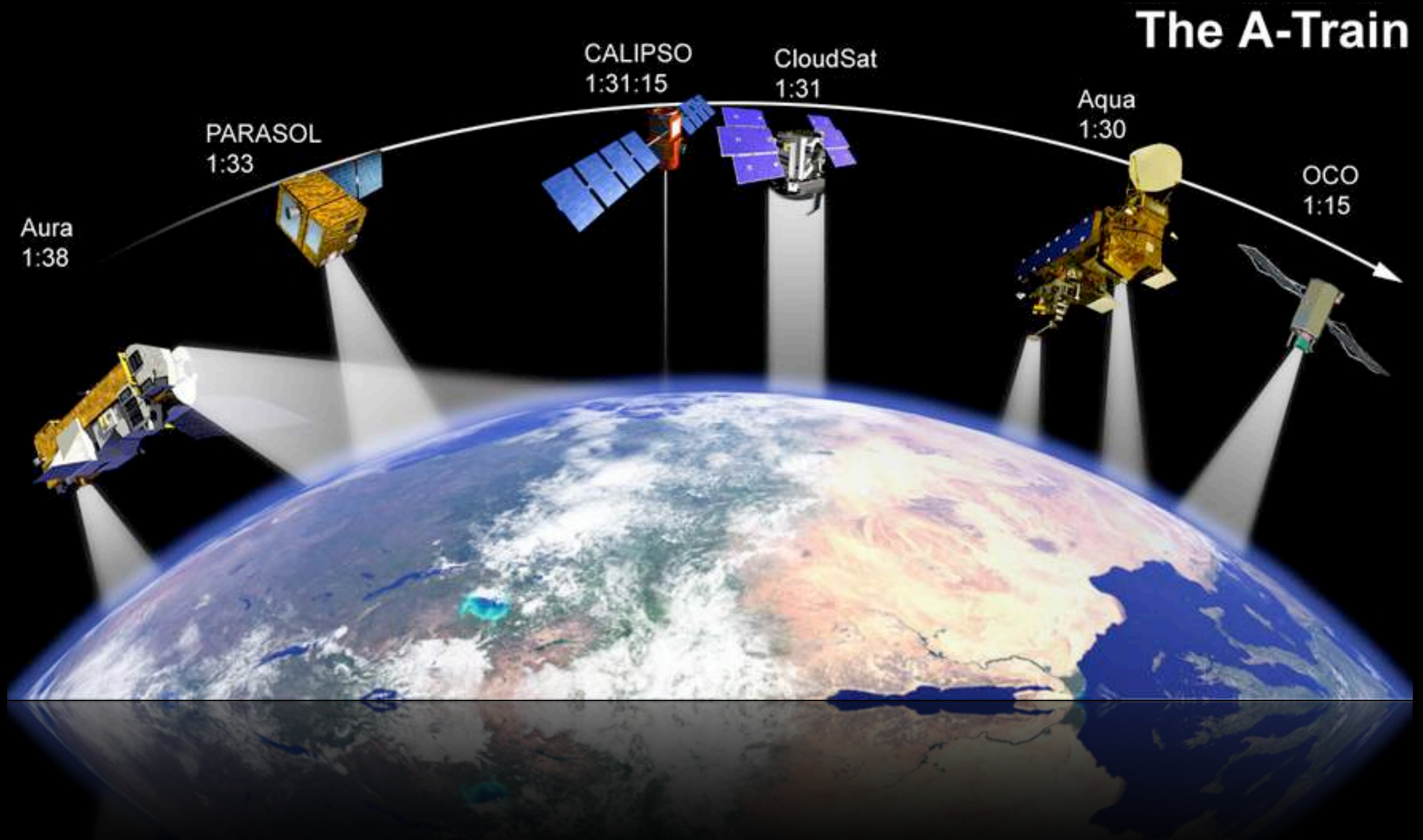


Hourly Measurements from 55 countries and more than 8,000 measurement sites from 1997-present





Sensing Assets





Sensing Assets

Clouds and Aerosols

Earth's water cycle

The A-Train

CALIPSO
1:31:15

CloudSat
1:31

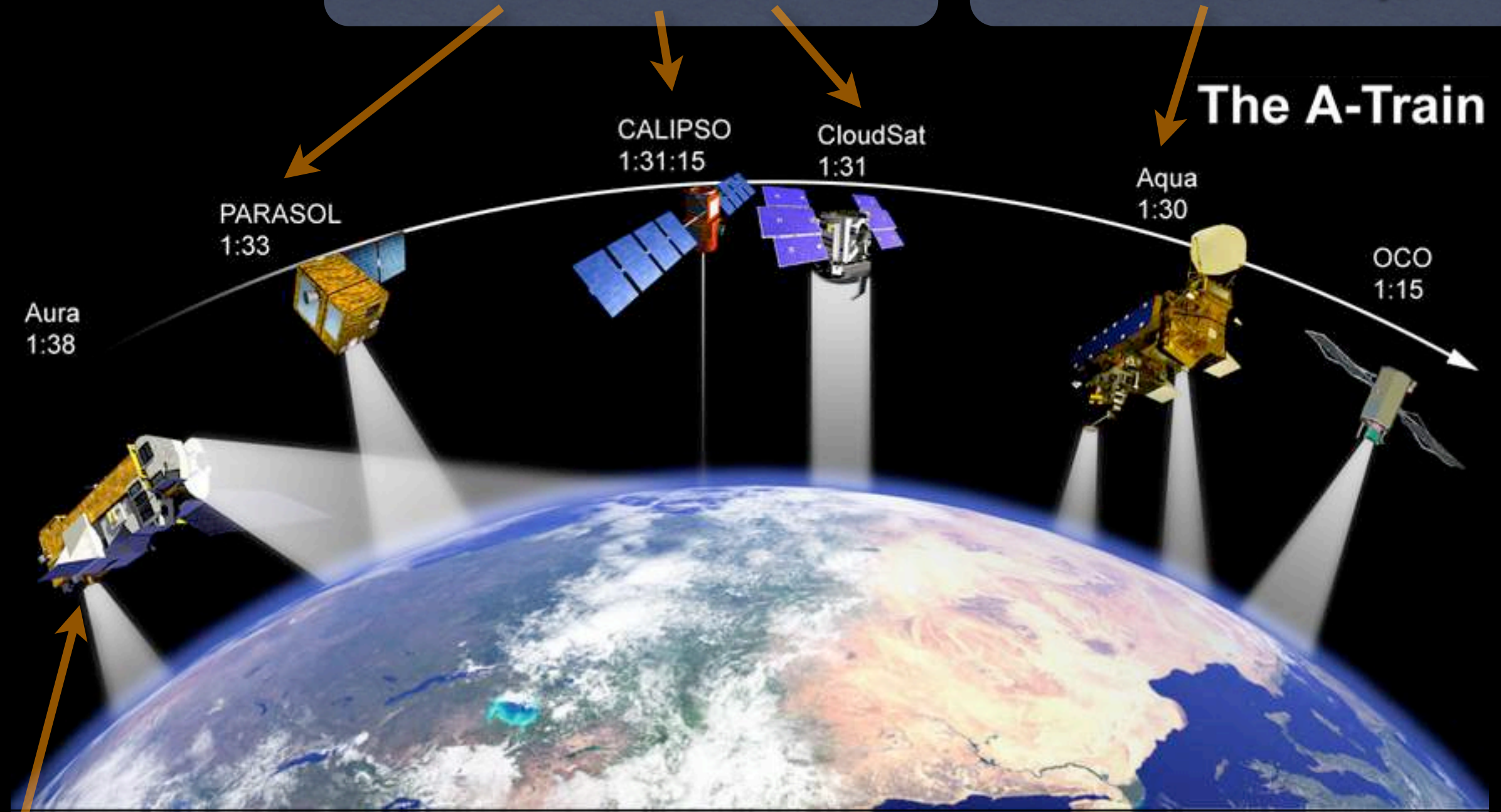
PARASOL
1:33

Aqua
1:30

OCO
1:15

Aura
1:38

Atmospheric Chemistry





Sensing Assets

Clouds and Aerosols

Earth's water cycle

The A-Train

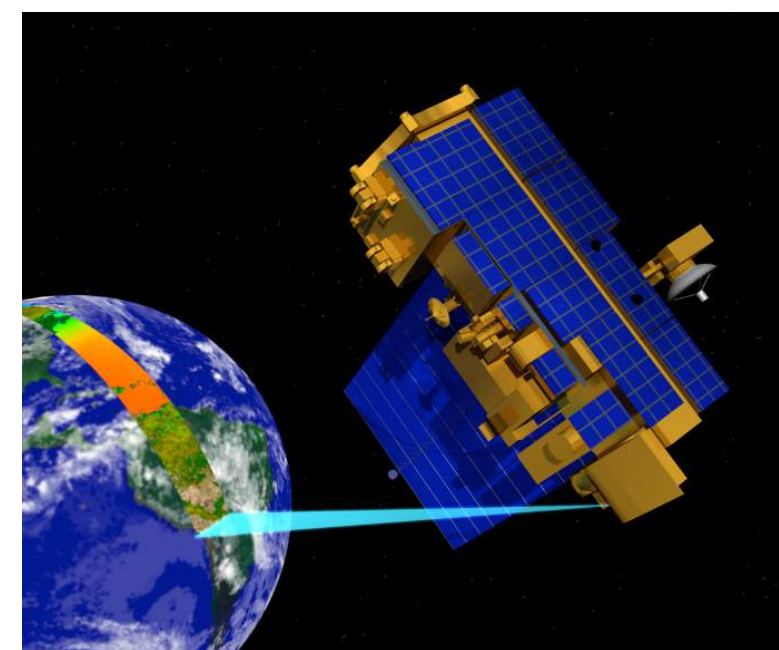
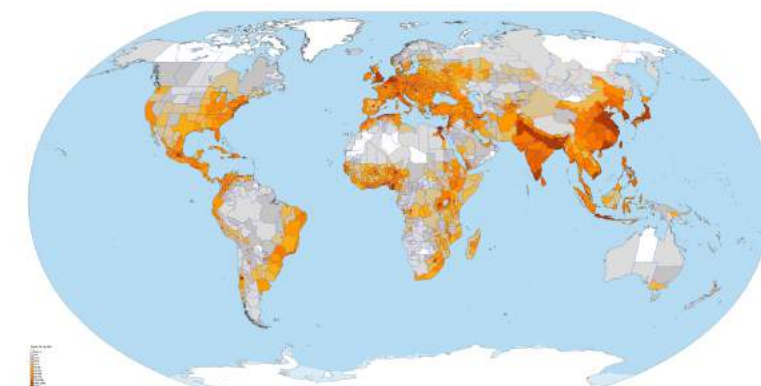


Atmospheric Chemistry



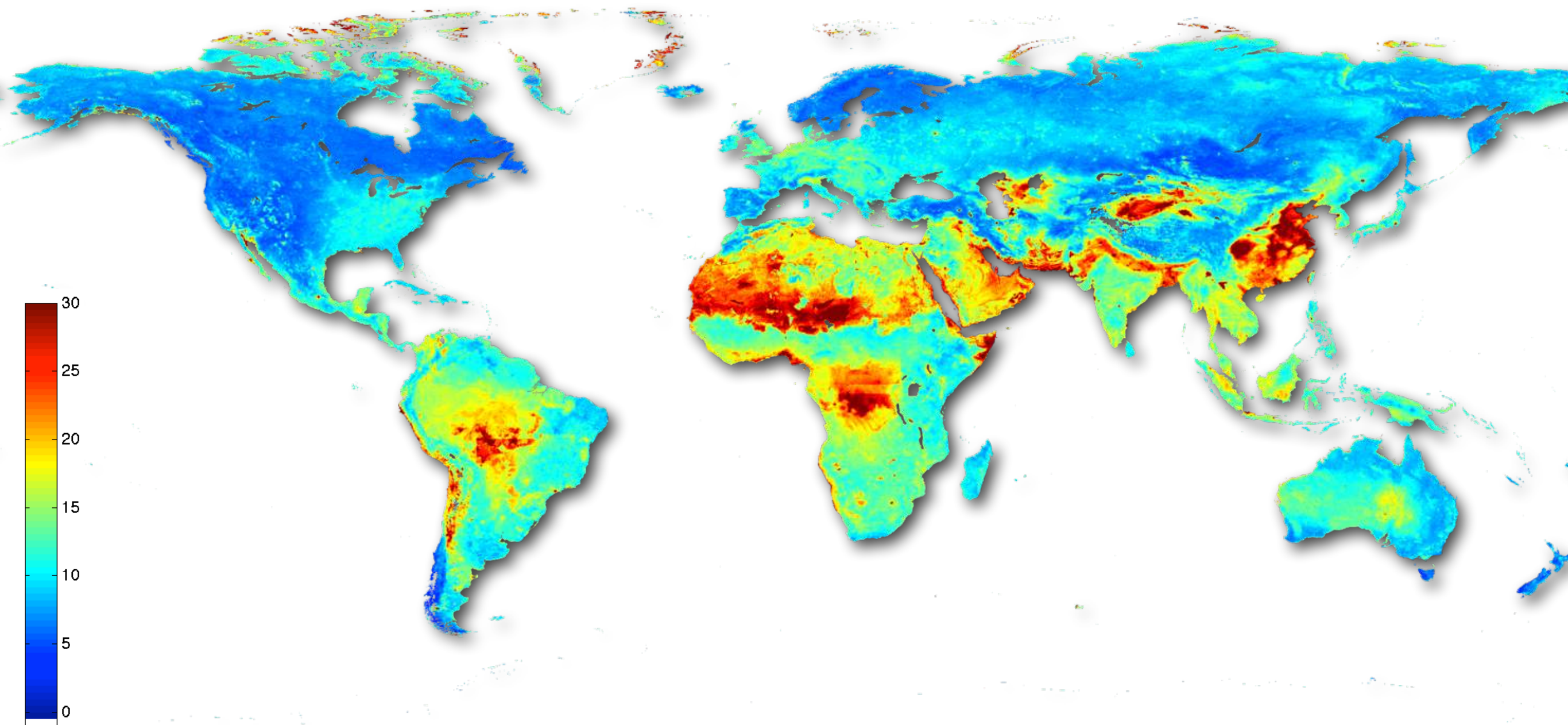
Aqua DeepBlue

Rank	Source	Variable	Type
1	Satellite Product	Tropospheric NO ₂ Column	Input
2	Satellite Product	Solar Azimuth	Input
3	Meteorological Analyses	Air Density at Surface	Input
4	Satellite Product	Sensor Zenith	Input
5	Satellite Product	White-sky Albedo at 470 nm	Input
6		Population Density	Input
7	Satellite Product	Deep Blue Surface Reflectance 470 nm	Input
8	Meteorological Analyses	Surface Air Temperature	Input
9	Meteorological Analyses	Surface Ventilation Velocity	Input
10	Meteorological Analyses	Surface Wind Speed	Input
11	Satellite Product	White-sky Albedo at 858 nm	Input
12	Satellite Product	White-sky Albedo at 2,130 nm	Input
13	Satellite Product	Solar Zenith	Input
14	Meteorological Analyses	Surface Layer Height	Input
15	Satellite Product	White-sky Albedo at 1,240 nm	Input
16	Satellite Product	Deep Blue Surface Reflectance 660 nm	Input
17	Satellite Product	Deep Blue Surface Reflectance 412 nm	Input
18	Satellite Product	White-sky Albedo at 1,640 nm	Input
19	Satellite Product	Sensor Azimuth	Input
20	Satellite Product	Scattering Angle	Input
21	Meteorological Analyses	Surface Velocity Scale	Input
22	Satellite Product	Cloud Mask Qa	Input
23	Satellite Product	White-sky Albedo at 555 nm	Input
24	Satellite Product	Deep Blue Aerosol Optical Depth 550 nm	Input
25	Satellite Product	Deep Blue Aerosol Optical Depth 660 nm	Input
26	Satellite Product	Deep Blue Aerosol Optical Depth 412 nm	Input
27	Meteorological Analyses	Total Precipitation	Input
28	Satellite Product	White-sky Albedo at 648 nm	Input
29	Satellite Product	Deep Blue Aerosol Optical Depth 470 nm	Input
30	Satellite Product	Deep Blue Angstrom Exponent Land	Input
31	Meteorological Analyses	Surface Specific Humidity	Input
32	Satellite Product	Cloud Fraction Land	Input
	In-situ Observation	PM_{2.5}	Target



Air Quality: Long-Term Average 1997-present

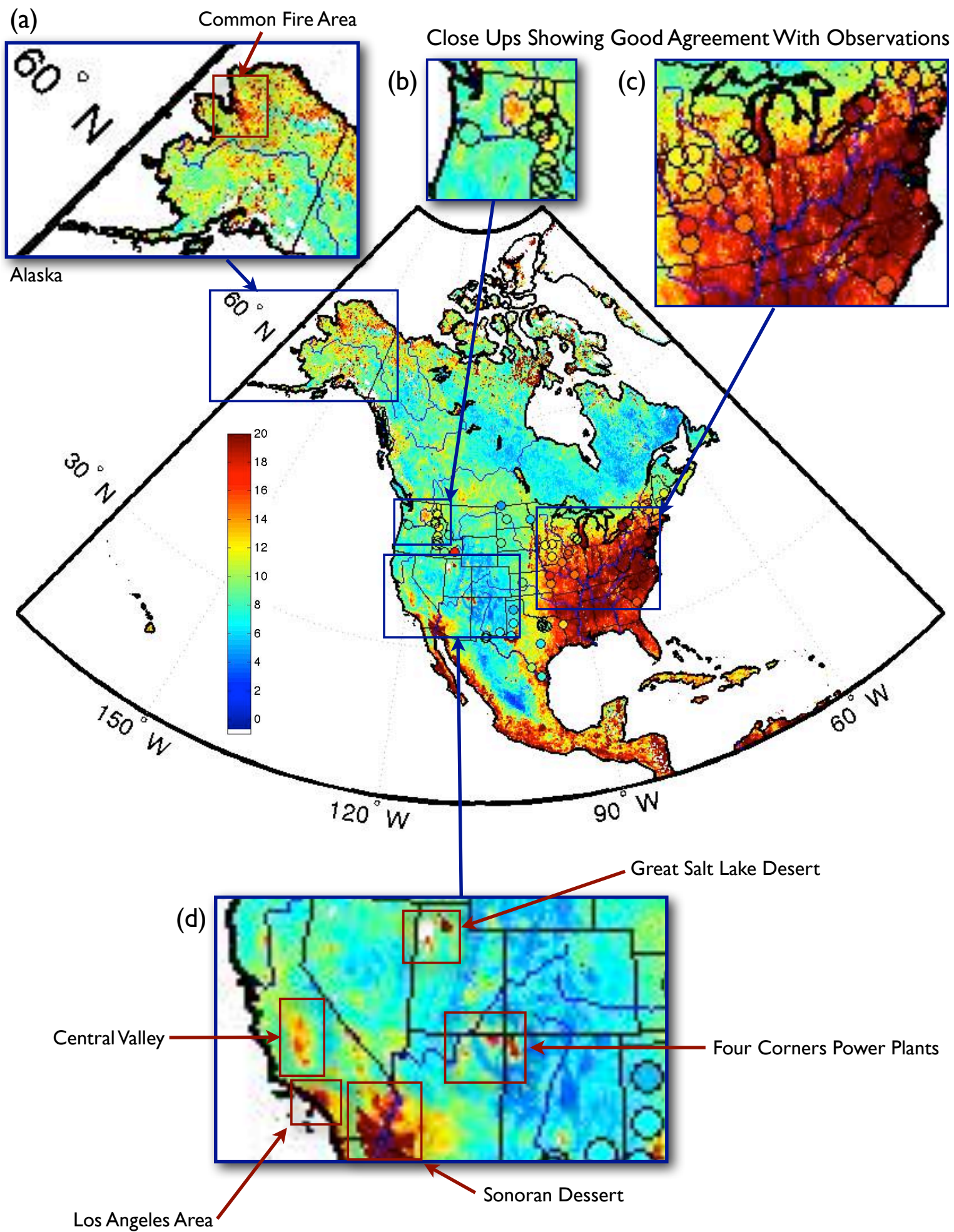
Used around 40 TB of different BigData sets from satellites, meteorology, demographics, in-situ sensors and scraped web-sites and social media to estimate PM_{2.5}.



This is a BigData Problem of Great Societal Relevance

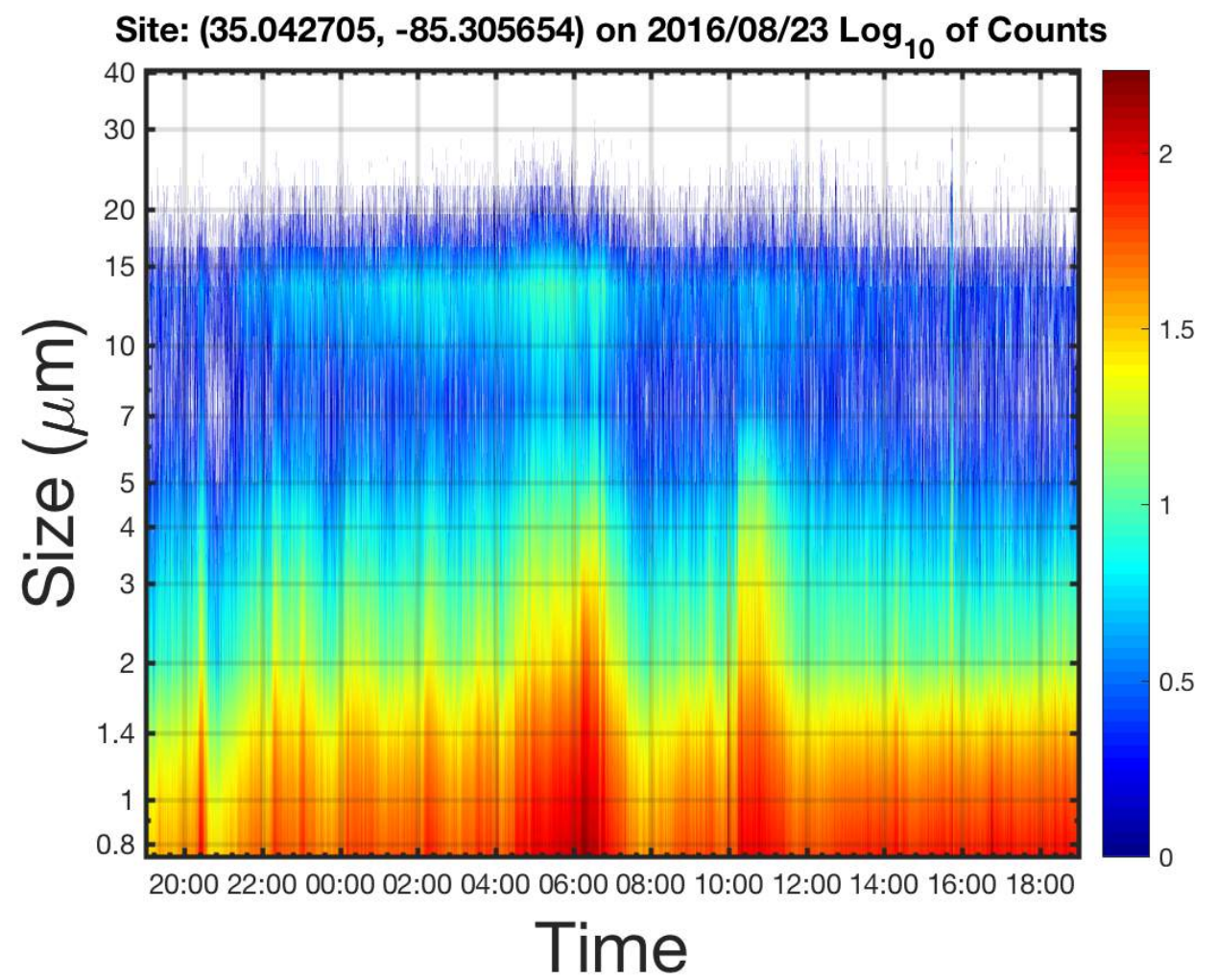
- Collecting data in real time from national and global networks requires **bandwidth**.
- With the next generation of wearable sensors and the **internet of things** this data volume will rapidly increase.
- A variety of applications enabled by **BigData**, **higher bandwidth** and **cloud processing**.
- Future finer granularity and **two way** communication will dramatically increase the size of the data bringing air quality to the micro scale, just like weather data.

	Time Taken			
	10 Mbps	20 Mbps	50 Mbps	1 Gbps
40 TB training data	185 days	93 days	37 days	1 day 21 hours
4 Gb update	54m	27m	11m	32s

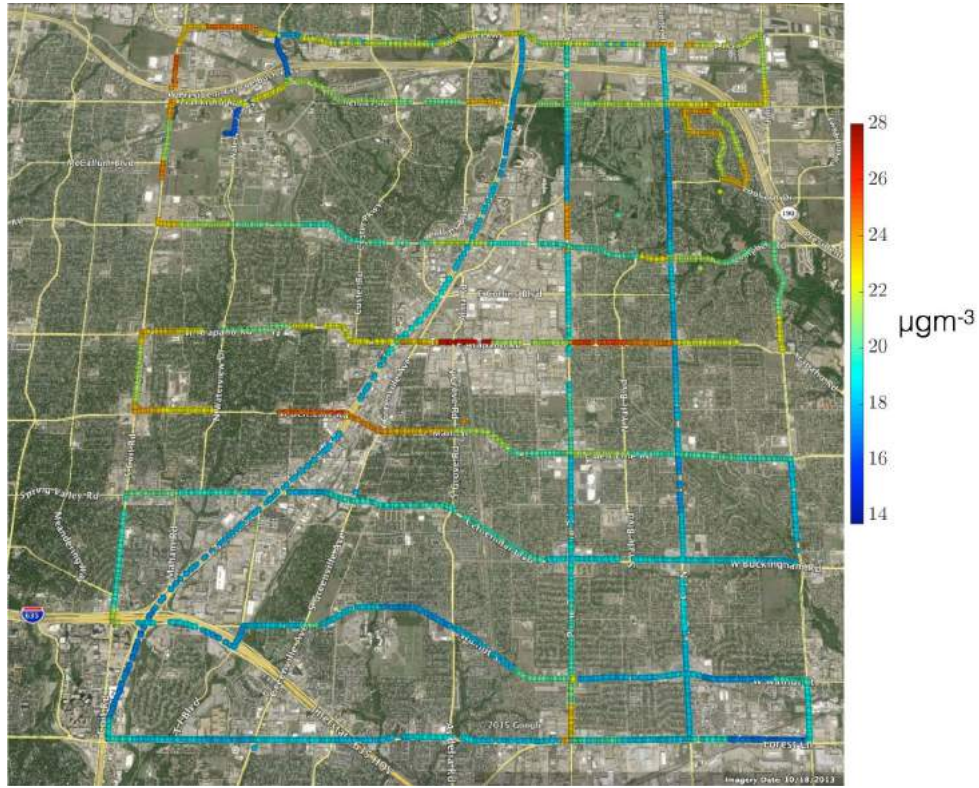




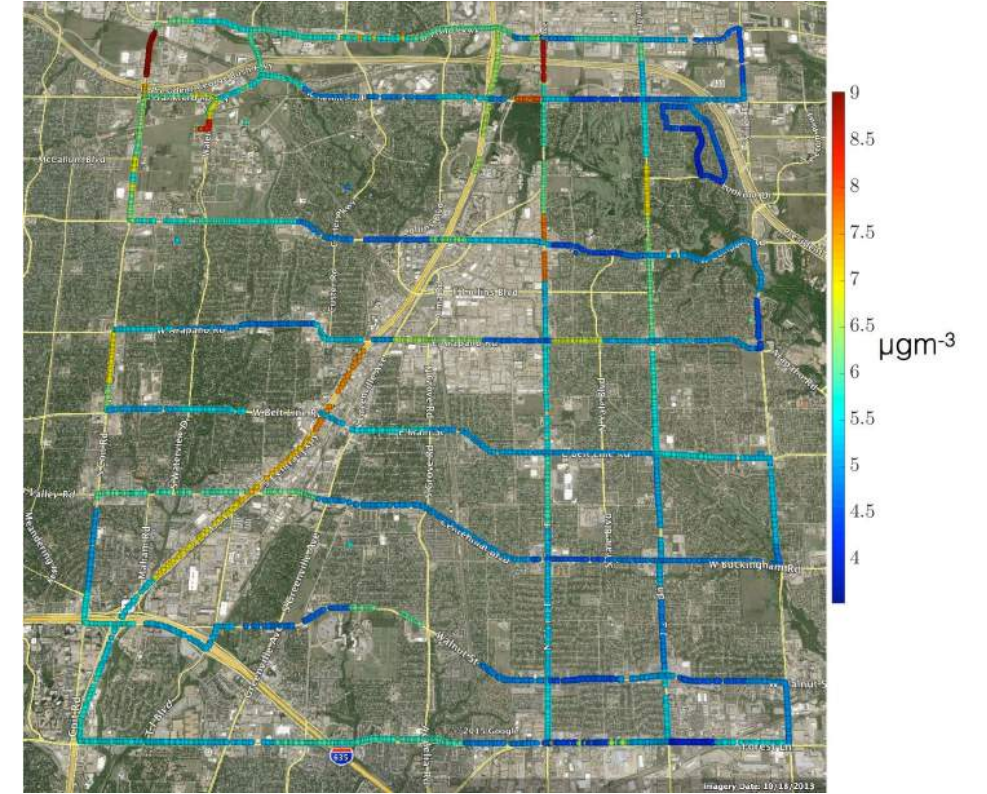
Ideal Spatial resolution is 0.5 km



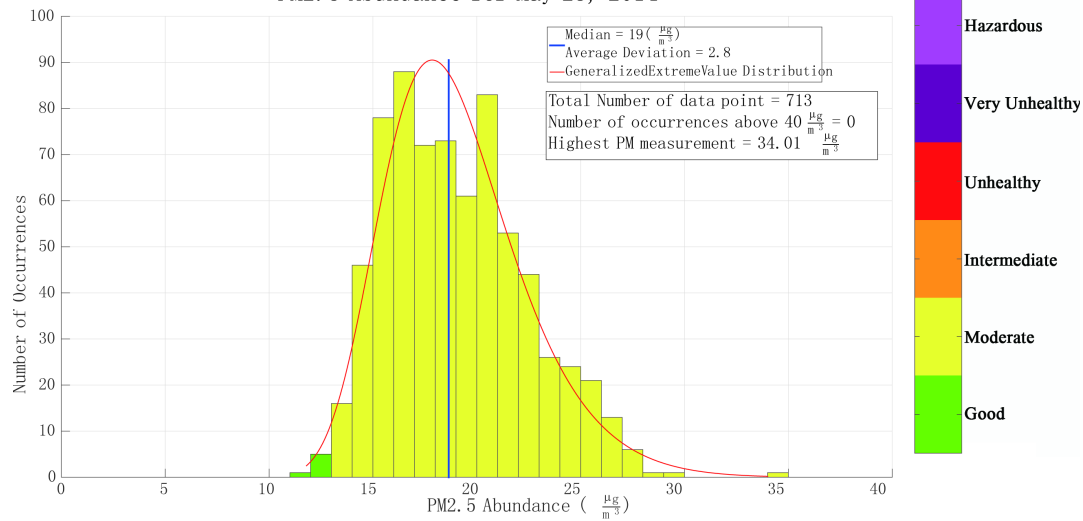
May 23, 2014



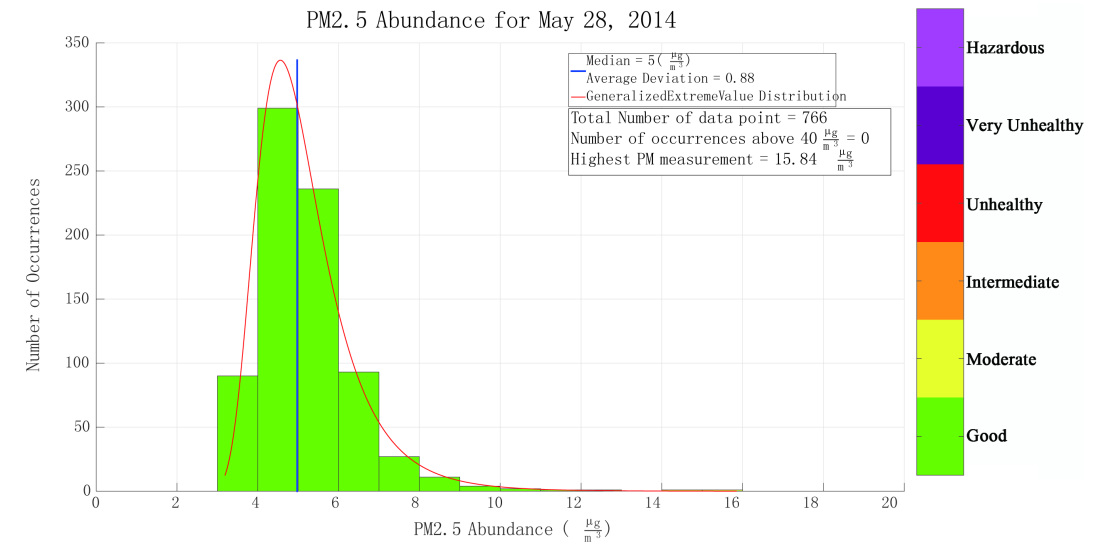
May 28, 2014



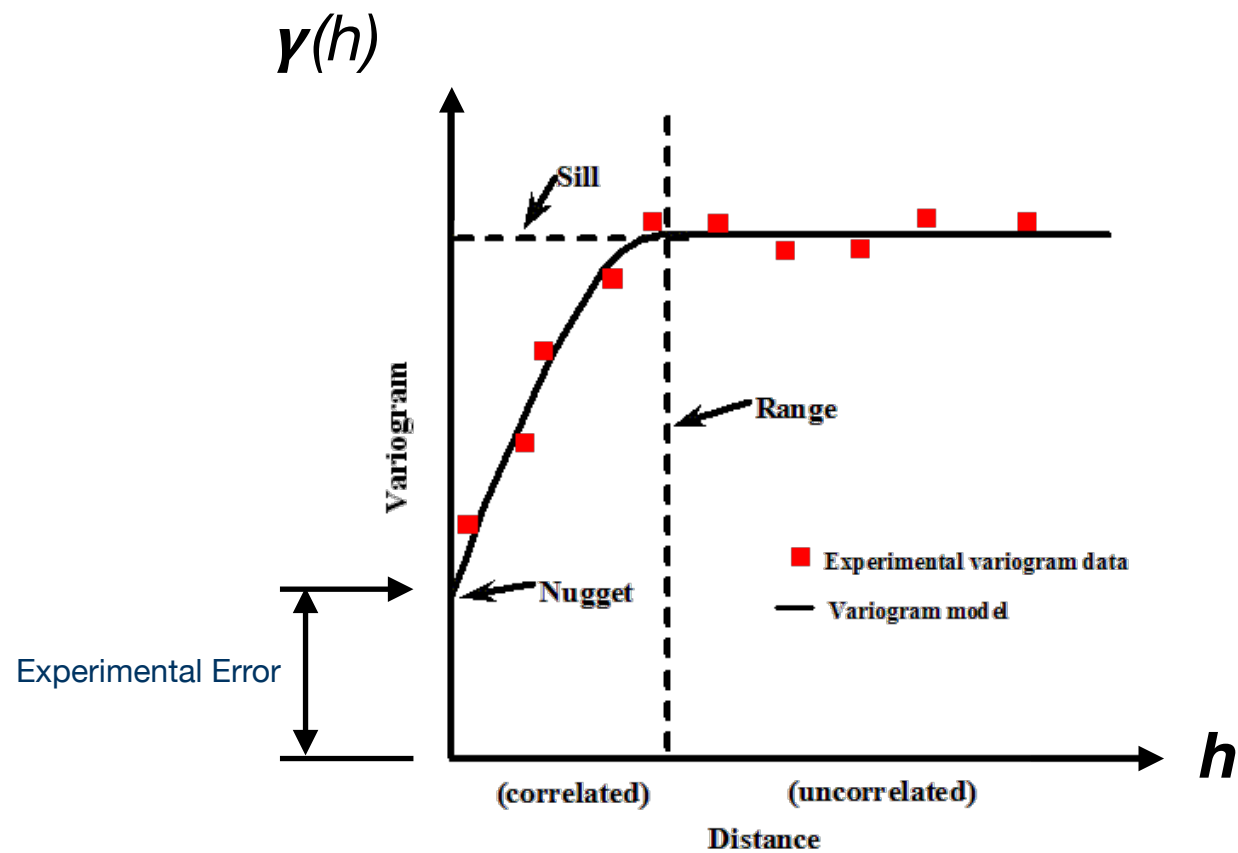
PM2.5 Abundance for May 23, 2014



PM2.5 Abundance for May 28, 2014



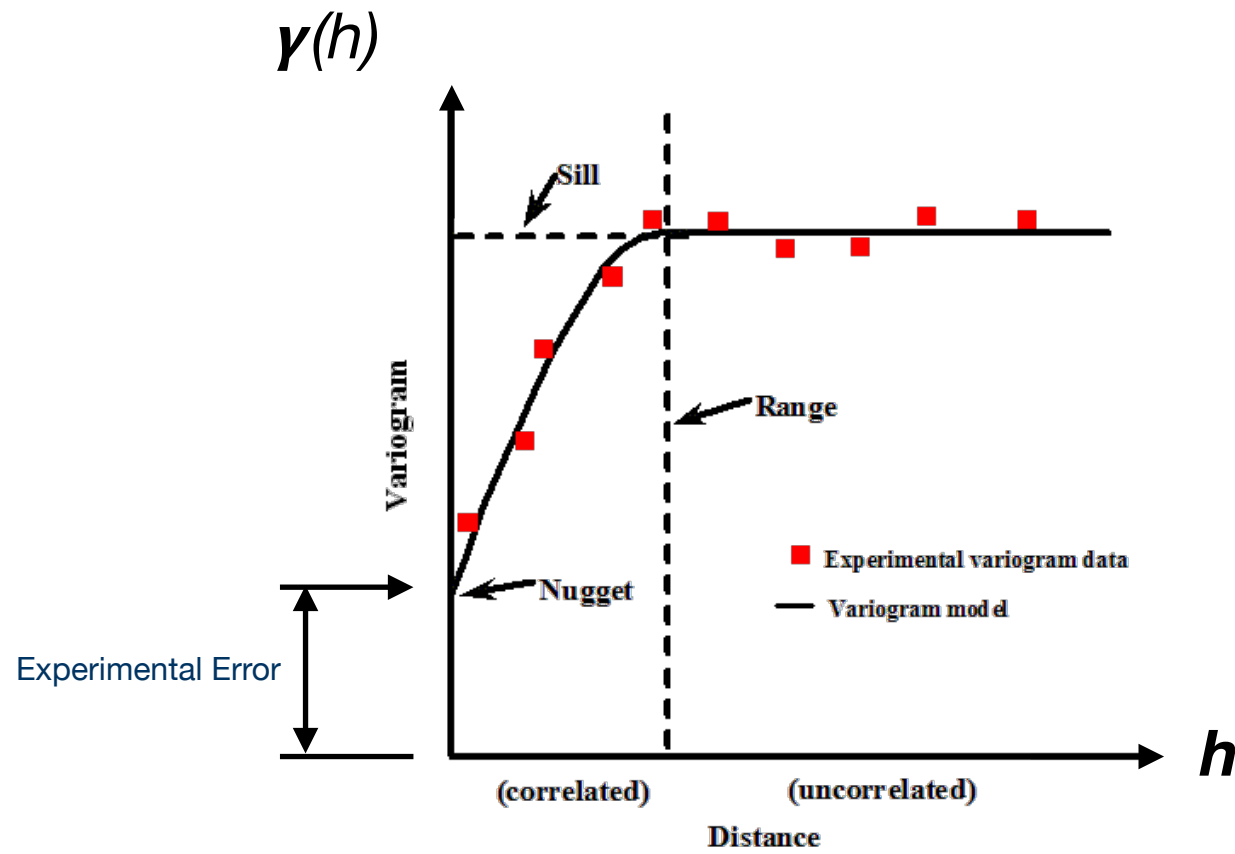
What Spatial Resolution?



$$\gamma(h) = \frac{\sum (y(x_i + h) - y(x_i))^2}{2N}$$

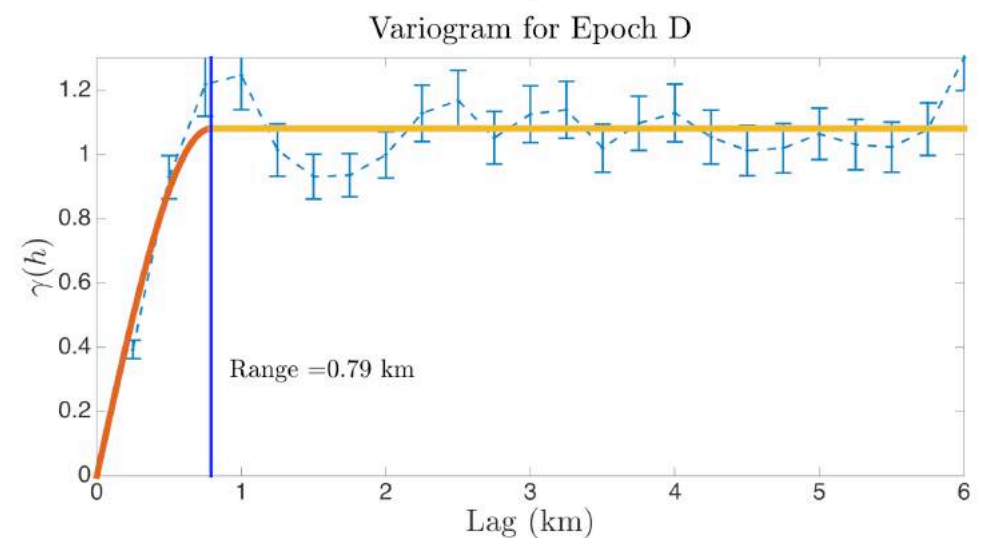
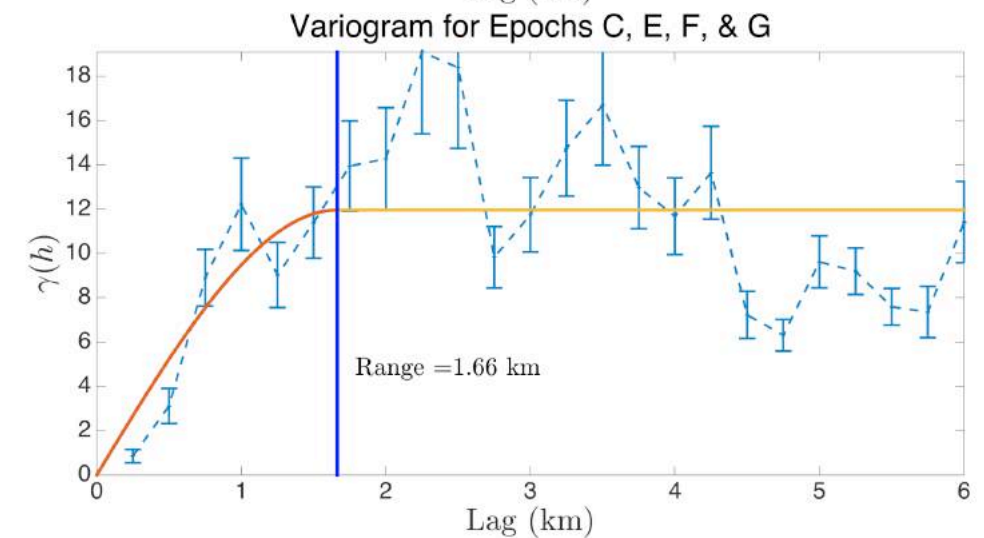
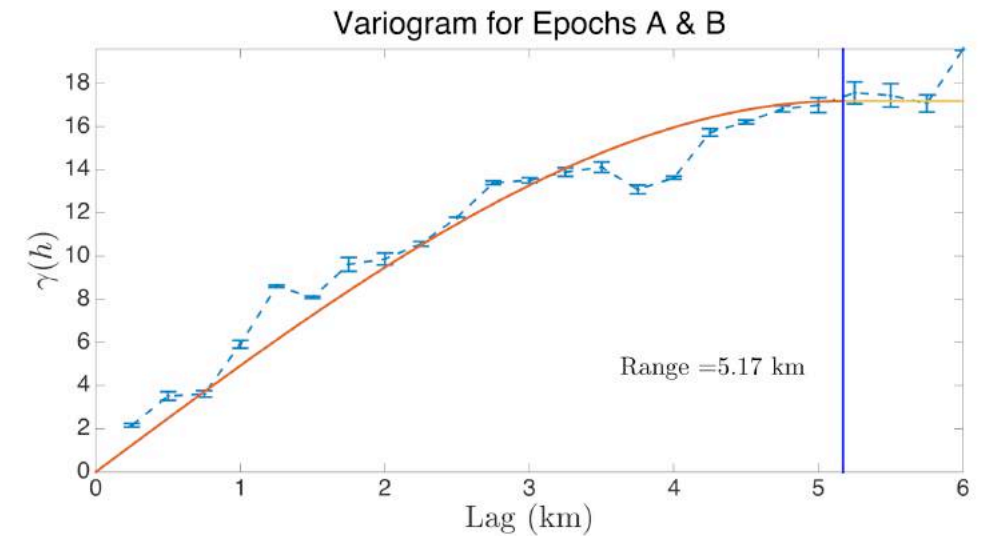
Half of the variance

What Spatial Resolution?



$$\gamma(h) = \frac{\sum (y(x_i + h) - y(x_i))^2}{2N}$$

Half of the variance



1



2

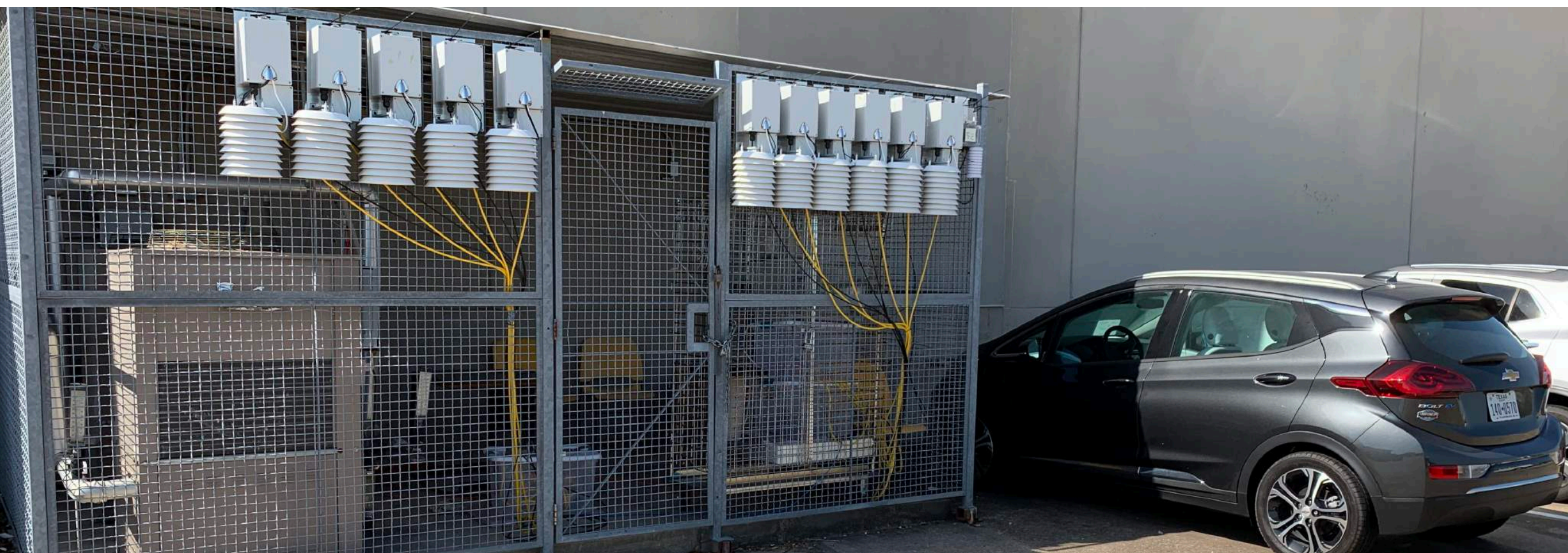


3



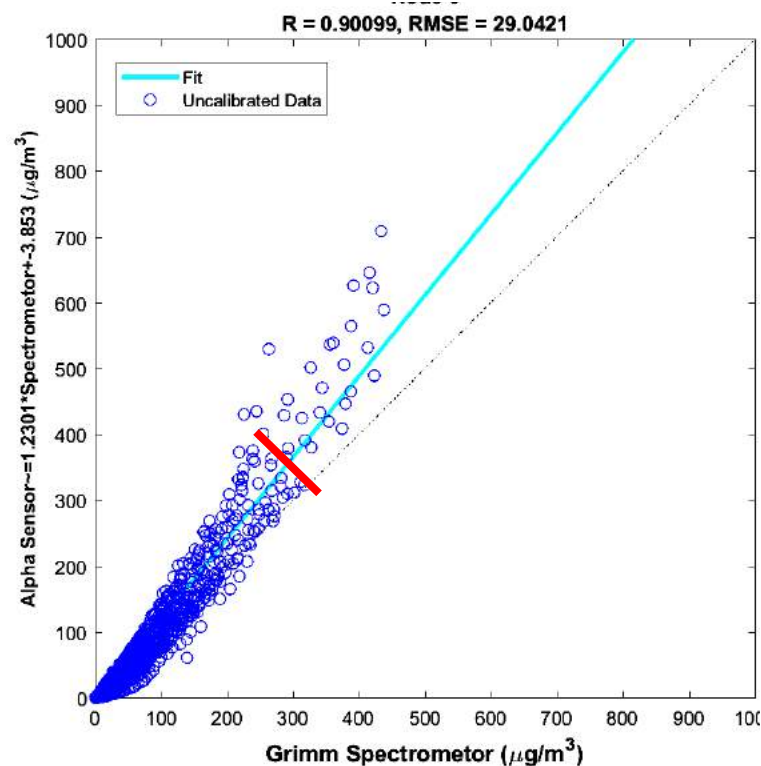
Cascade of accuracies

- 1. EPA certified instrument: \$25,000-\$50,000 (primary)
- 2. Medium accuracy: \$2,000-\$5,000 (secondary)
- 3. Inexpensive but useful: \$200-\$500 (tertiary)

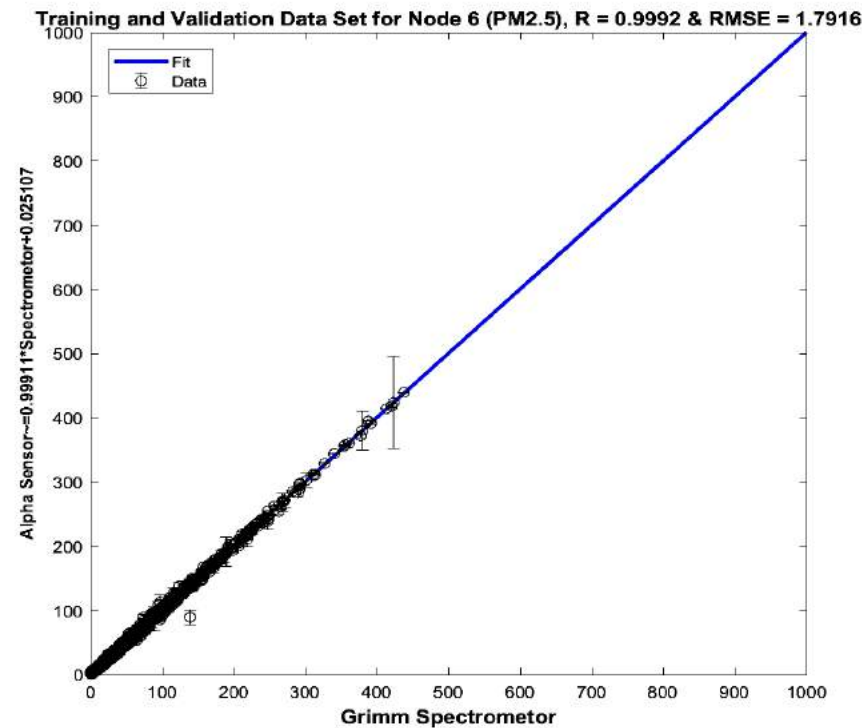


Example Machine Learning Calibration

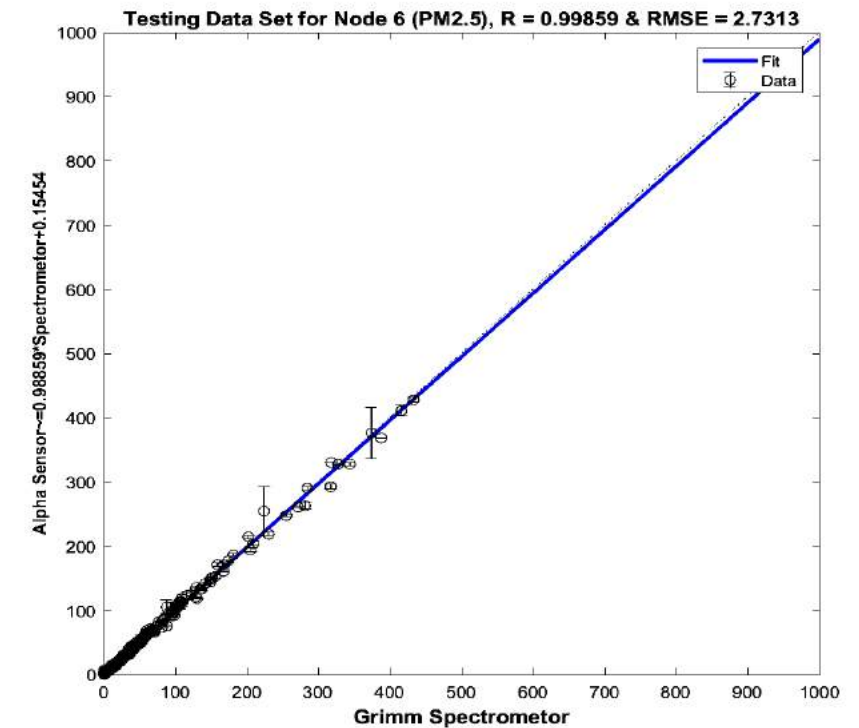
Calibration is greatly improved when it is multivariate, nonlinear and parametric.



Uncalibrated
Data Set



Training Data Set

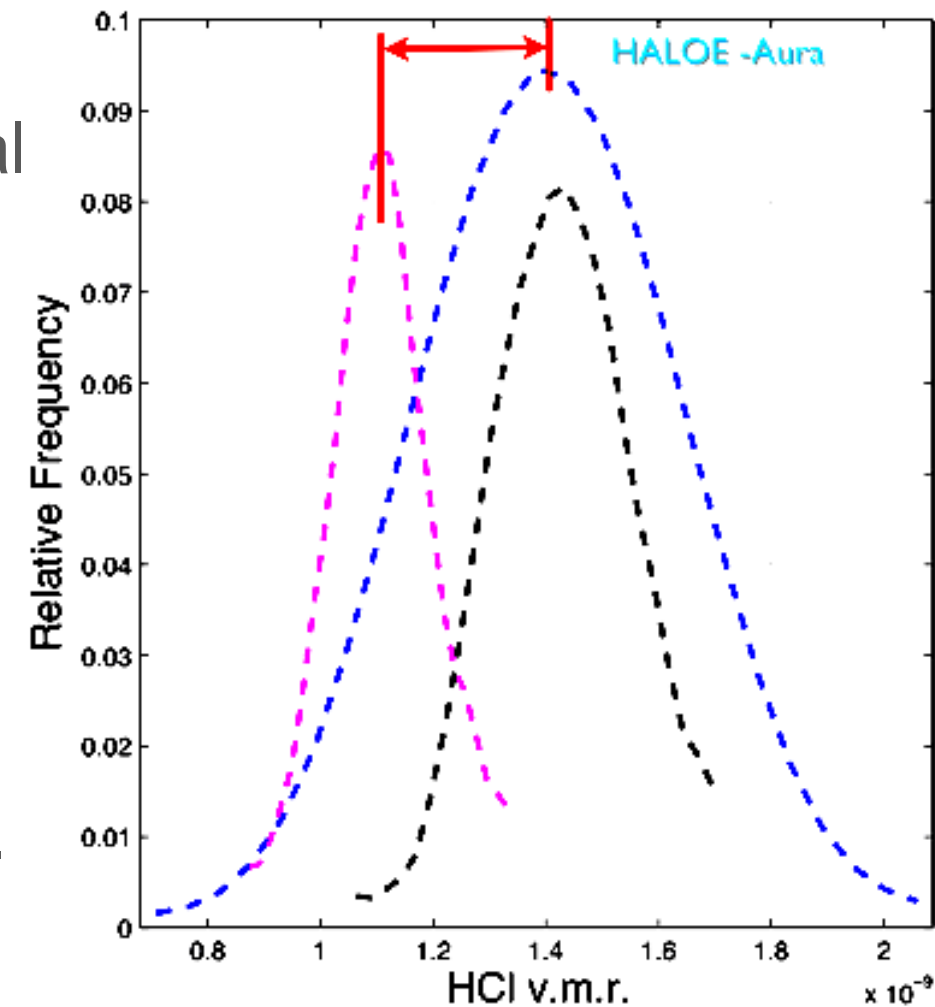


Independent
Validation
Note the inclusion
of error estimates.

Representativeness

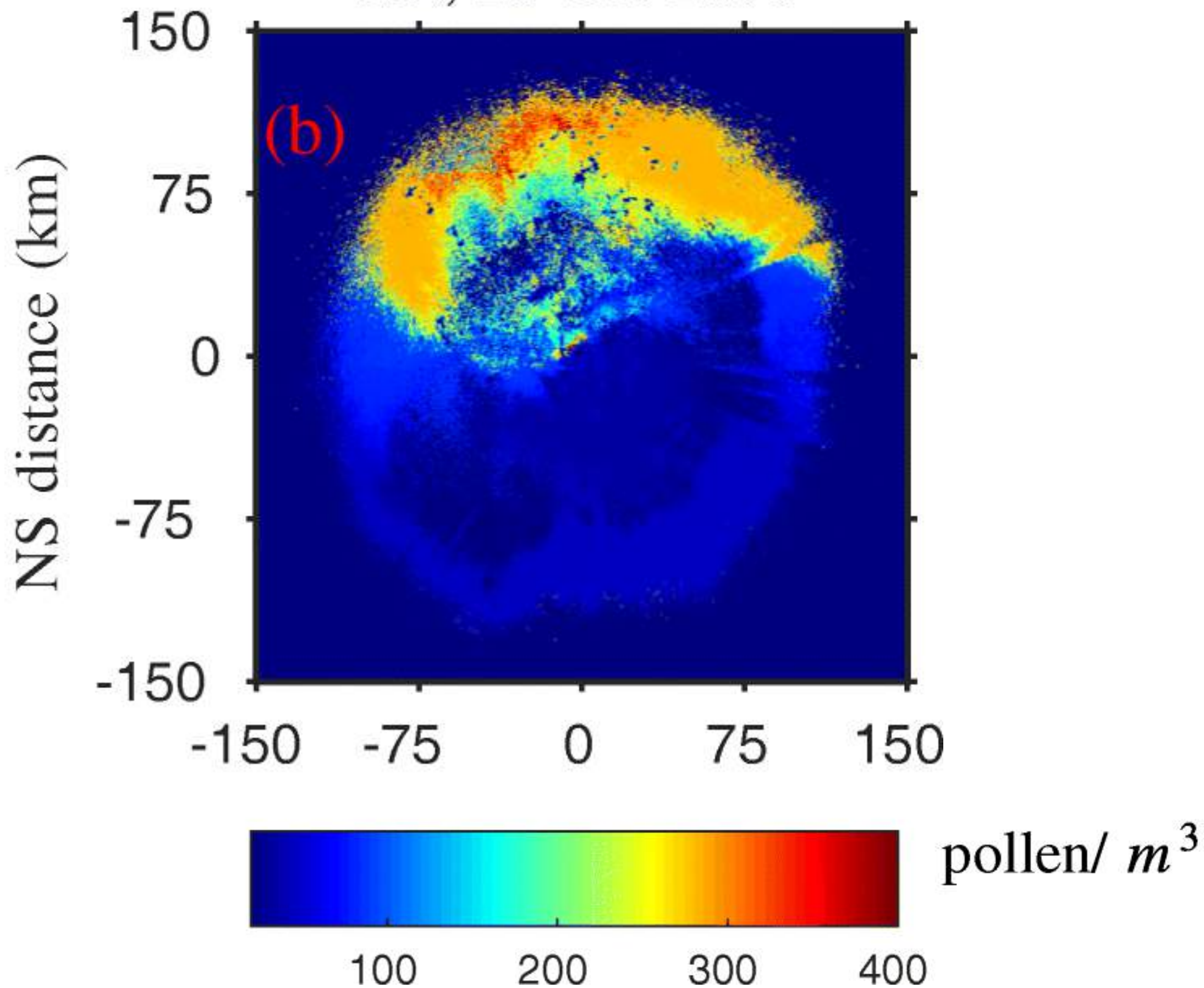
- When performing chemical data assimilation the observational, representativeness, and theoretical uncertainties have very different characteristics.
- We routinely accurately characterize the representativeness uncertainty by studying the probability distribution function (PDF) of observations. The average deviation has been used as a **measure of the width** of the PDF and of the variability (representativeness uncertainty).
- The representativeness uncertainty can be markedly different from the observational uncertainty and clearly delineates mixing barriers.

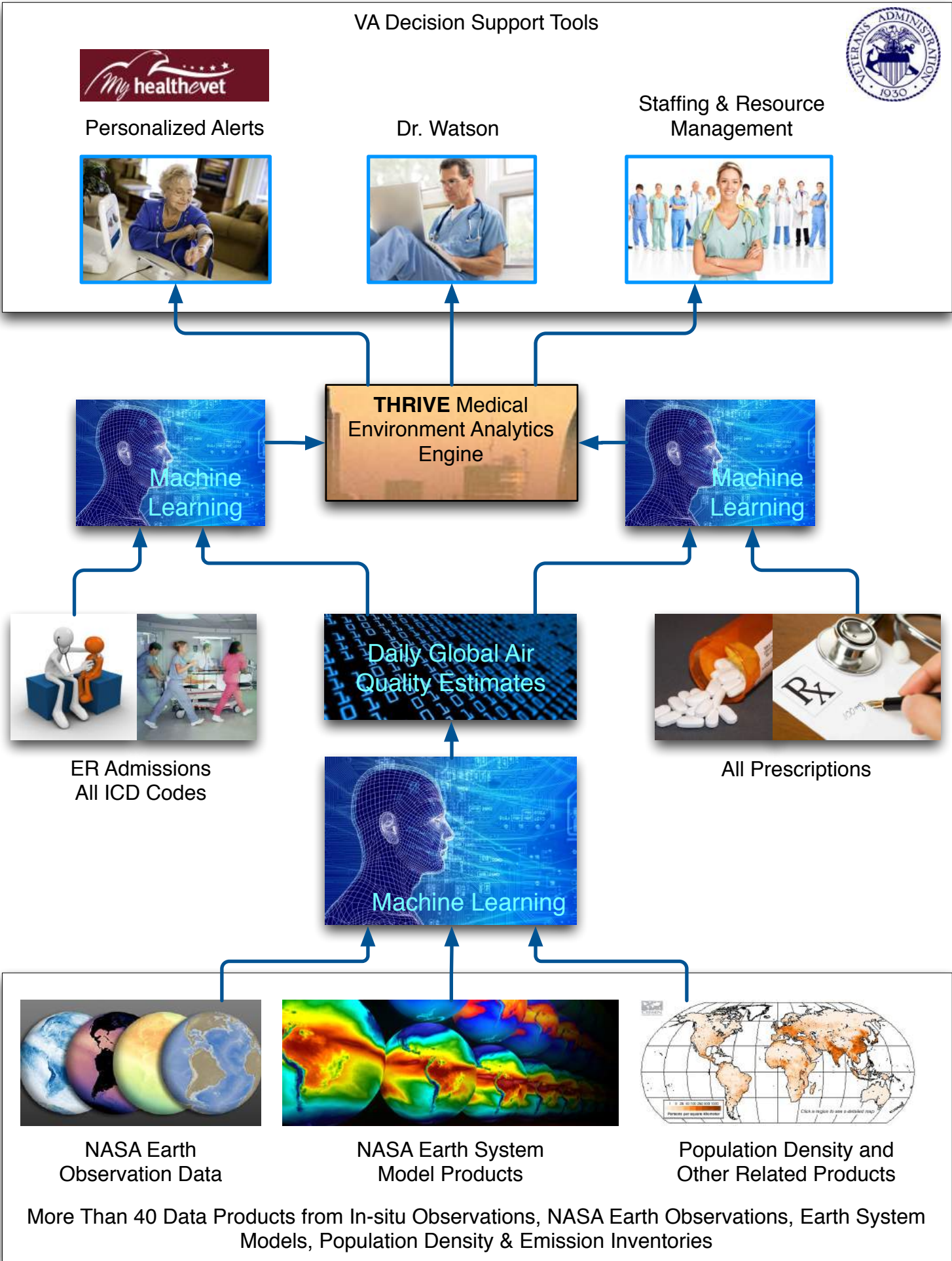
Source: doi:10.1016/
j.atmoscilet.2003.11.002



$$\sigma_{\text{rep}} = \text{ADev}(\chi_1, \dots, \chi_N) = \frac{1}{N} \sum_{j=1}^N |\chi_j - \bar{\chi}|$$

NN, 28-Oct-2008





Biometrics & Environmental Measurements

- IRB approval granted
 - Environmental Sensors
 - Biometric Sensors
- Building/calibrating 130 sensor network.
- Machine learning tools are in place.
- Subject recruiting about to start.



CAP



SNAP/DIN



32



Biometric Measurements



PPG/SpO2/HR
Oxygen saturation

Pupil Dilation
Cognitive Load

Respiration

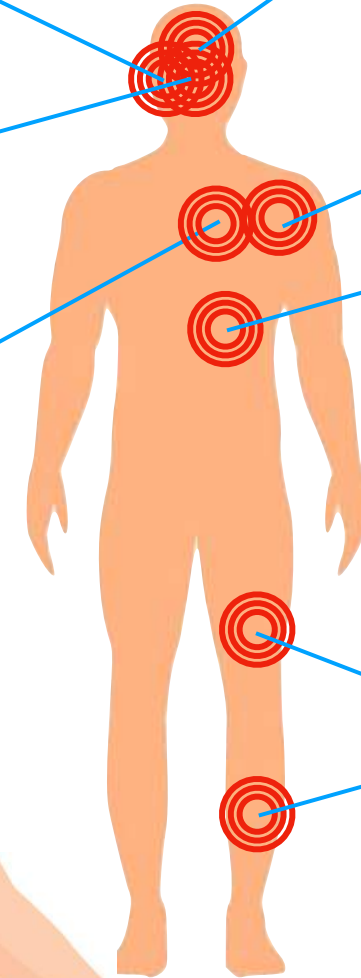
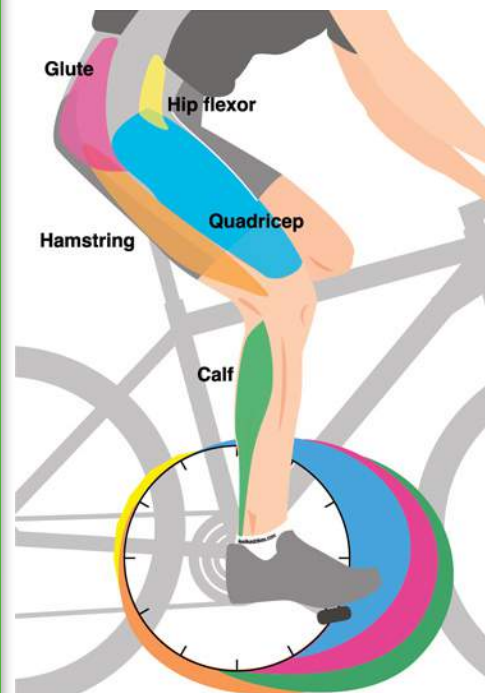
EEG
Brain Activity
e.g. Cognitive Processes

GSR
Skin Conductance
e.g. Alertness

ECG
Heart Activity
e.g. Anxiety

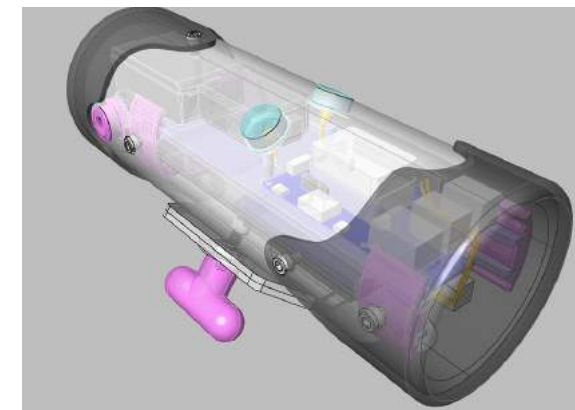
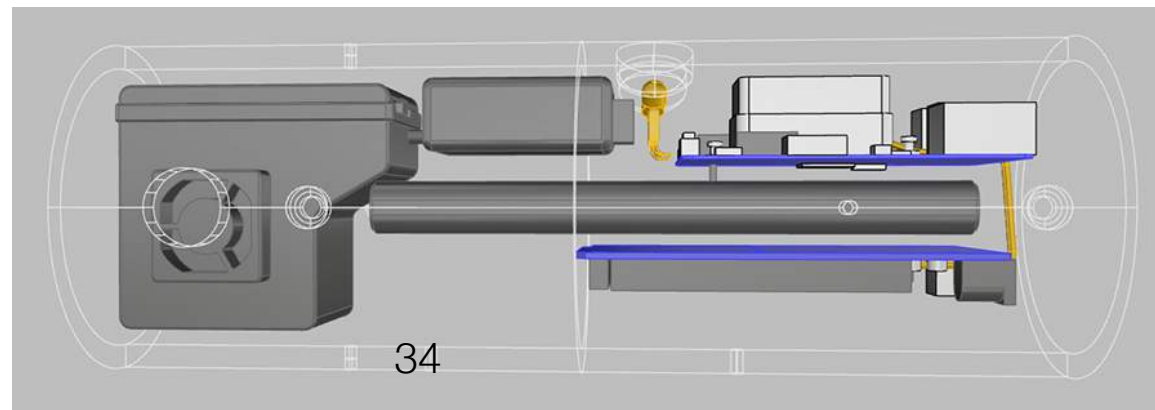
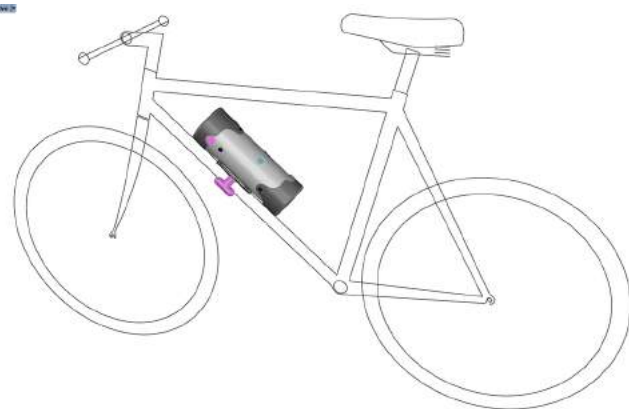
EMG
Muscle Activity
e.g. Performance

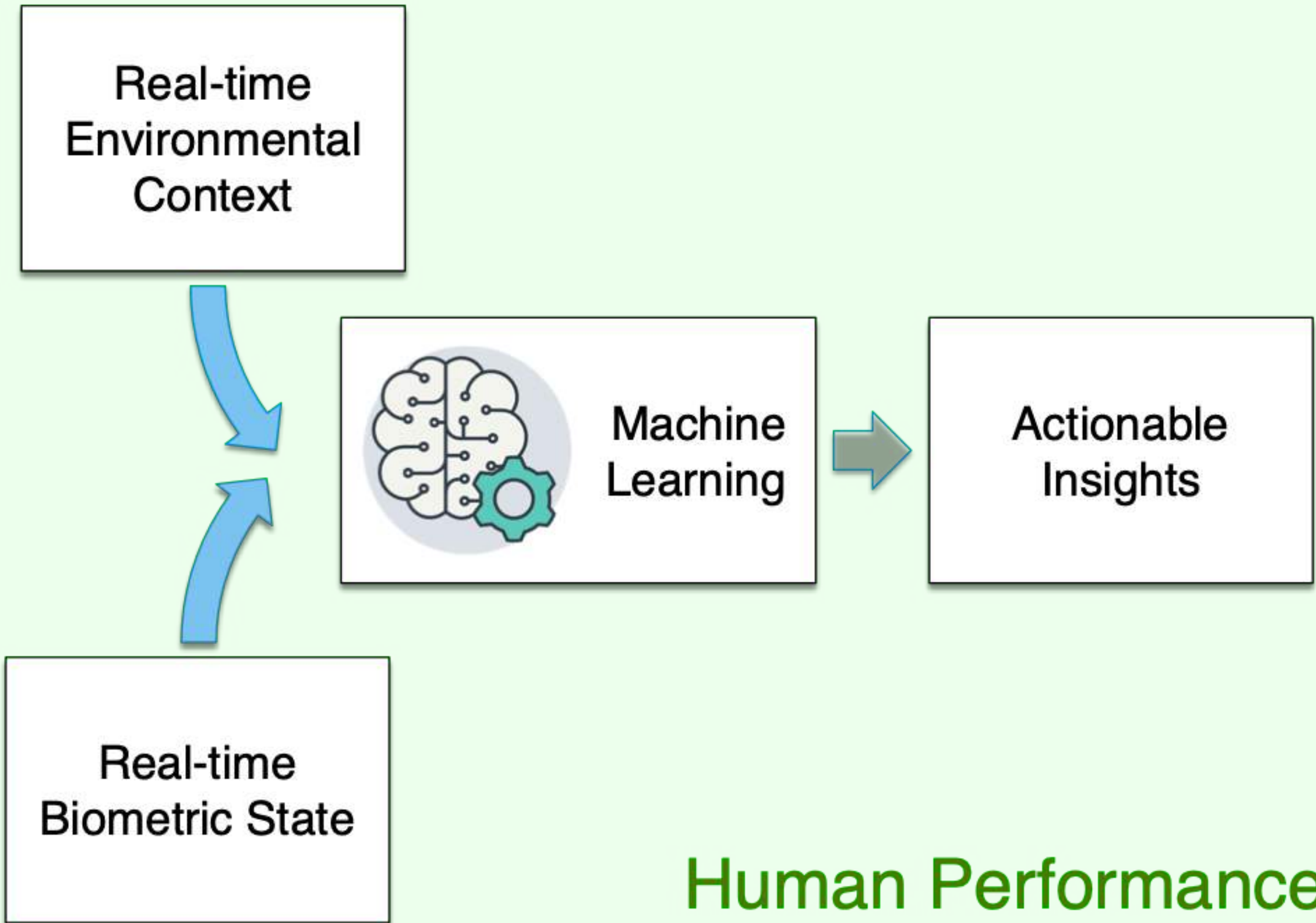
and more
Temperature,
HRV, IMU,

Environmental Measurements

- Particle spectrometers measuring from 10 nm - 100 microns in a few hundred size bins.
- Mass spectrometer measuring molecules up to 300 amu. With inlets at top and bottom of the car so we can measure vertical gradient in heavier than air gasses.
- NIST calibrated illuminance Spectrophotometer (360-780 nm).
- Distributed sensors streaming 24/7.

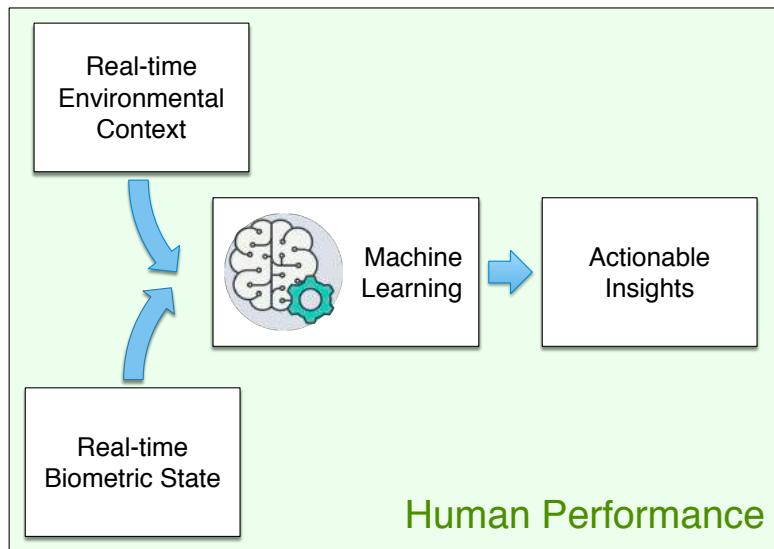




MINTS Comprehensive Sensing & Machine Learning

Multi-scale Integrated Interactive Intelligent Sensing and Simulation

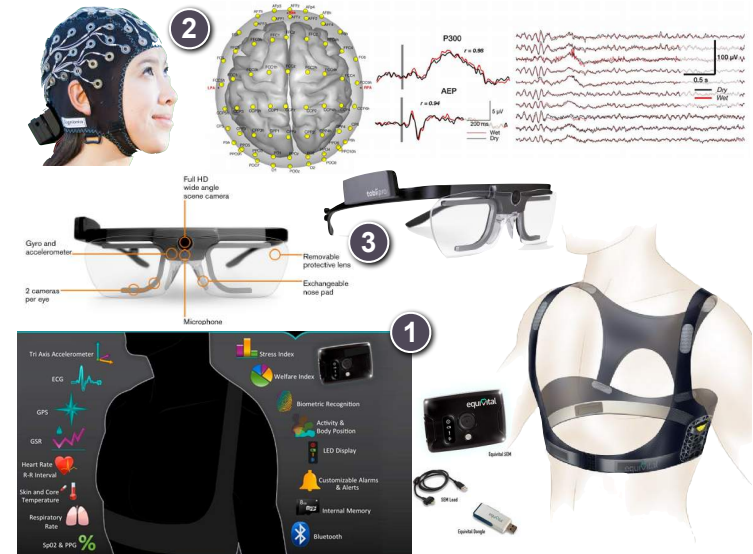
CBRN (Chemical Biological Radiological Nuclear) Sentinels For Actionable Insights



MINTS Comprehensive Context Engine

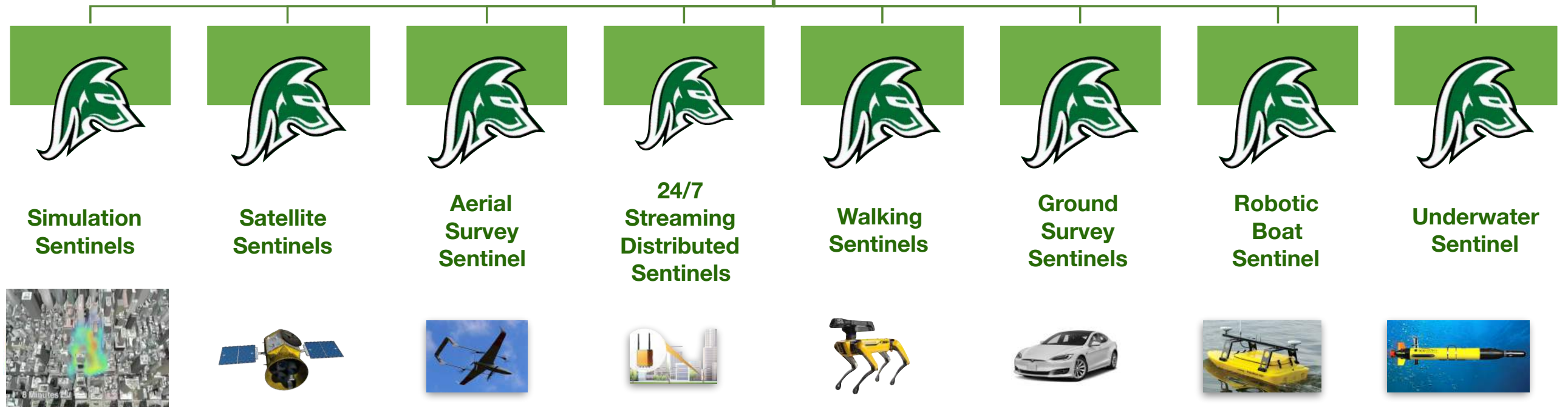


Biometrics Package

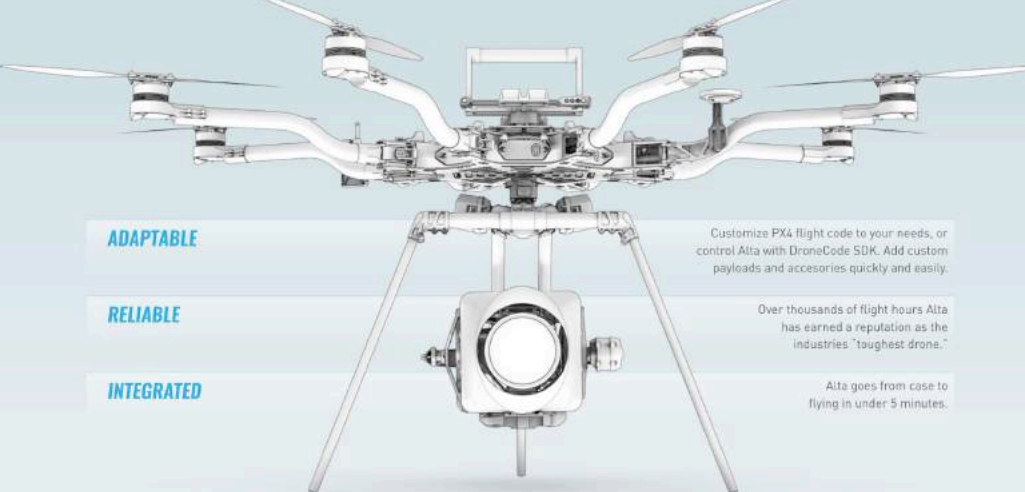


Schematic showing the holistic biometric sensing environment we propose making the human response an integral part of the sensor network. (1) Equival Black Ghost system, (2) Cognionics 64 electrode EEG cap, and (3) Tobii Pro Glasses 2 for eye tracking.

Eight Sentinel Types







ADAPTABLE Customize PX4 flight code to your needs, or control Alta with DroneCode SDK. Add custom payloads and accessories quickly and easily.

RELIABLE Over thousands of flight hours Alta has earned a reputation as the industries "toughest drone."

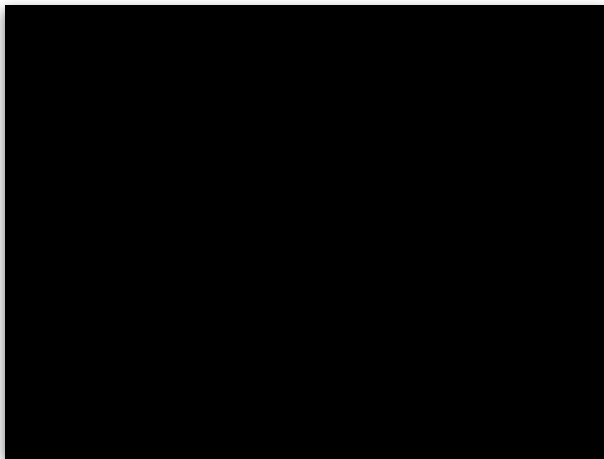
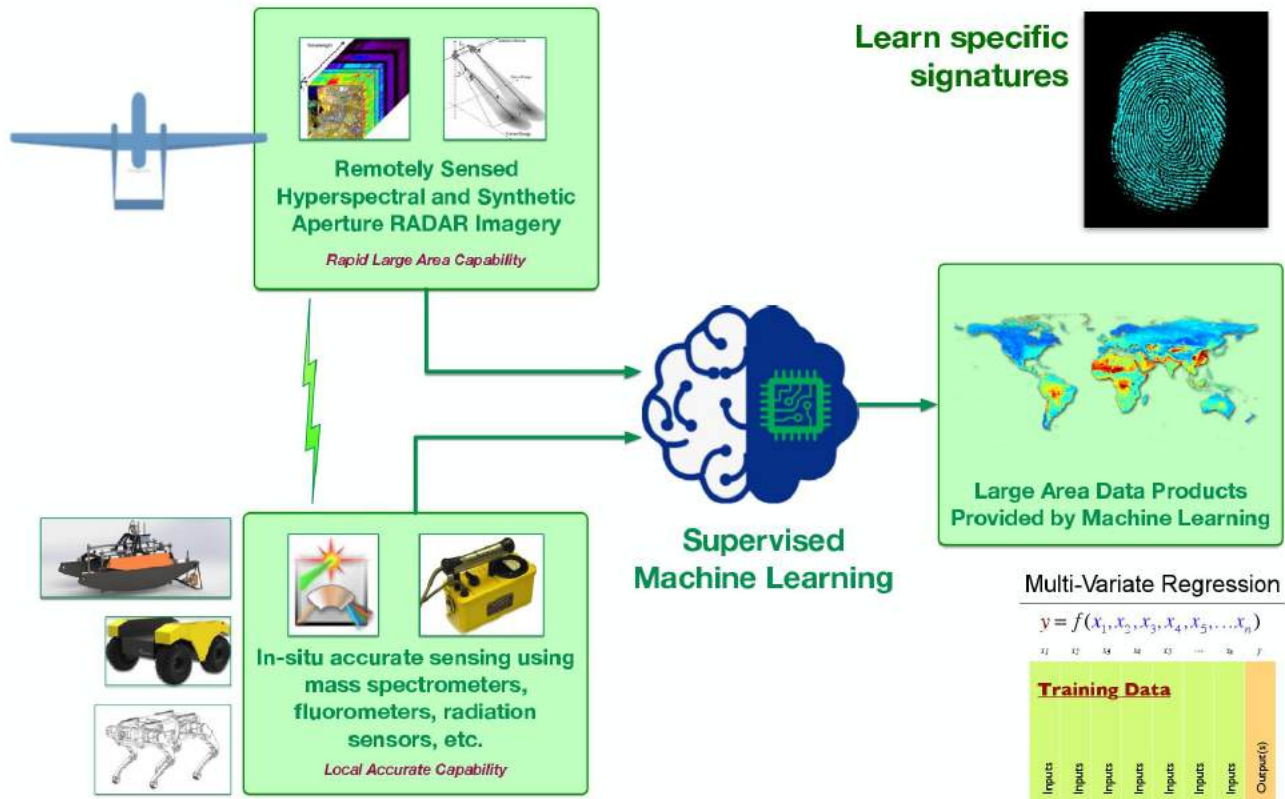
INTEGRATED Alta goes from case to flying in under 5 minutes.

FLIGHT MODES
MANUAL / HEIGHT MODE / POSITION MODE / RETURN-TO-LAND (RTL) / WAYPOINT MISSION MODE

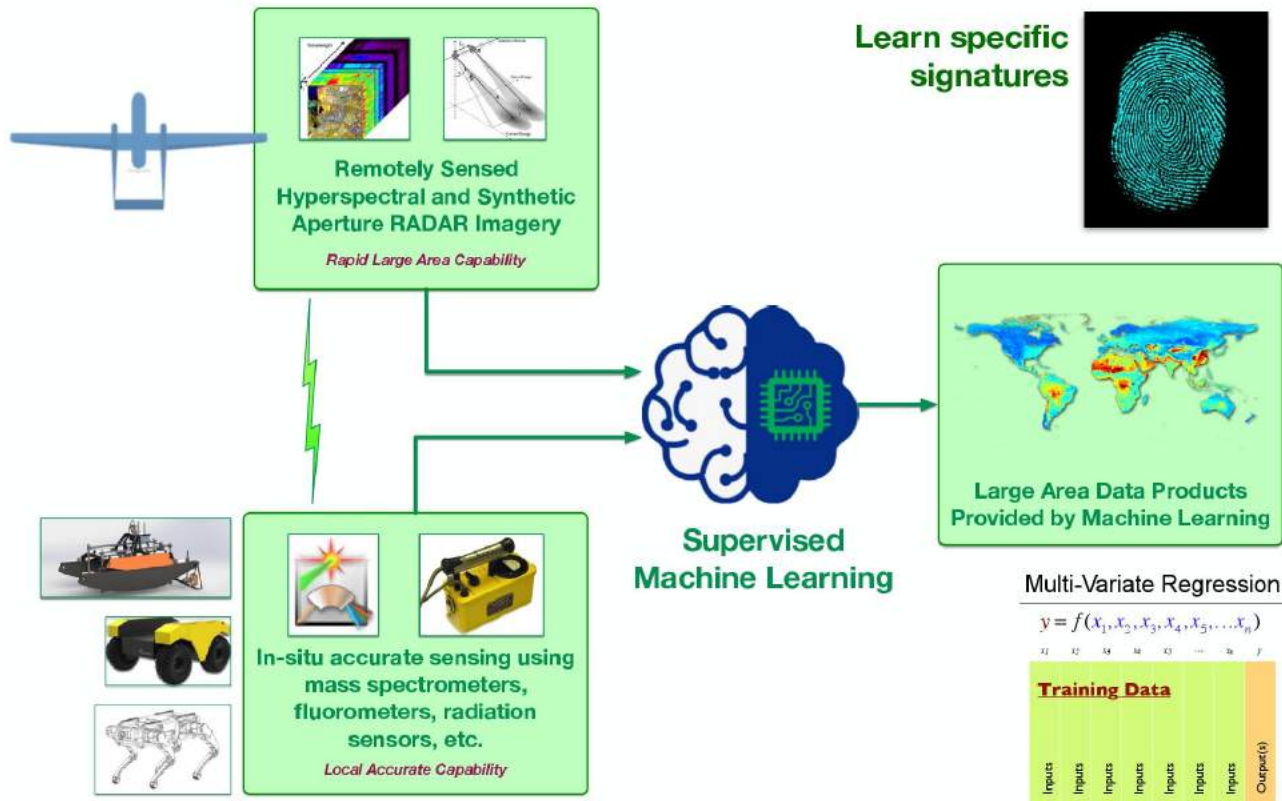
13.6 <small>lbs</small> WEIGHT	20 <small>lbs</small> MAXIMUM PAYLOAD	1325 <small>mm</small> UNFOLDED DIAMETER	660 <small>mm</small> FOLDED DIAMETER	145 <small>W/kg</small> TYPICAL SPECIFIC POWER	1.85 : 1 THRUST RATIO (AT MAX WEIGHT)
--	---	--	---	--	---



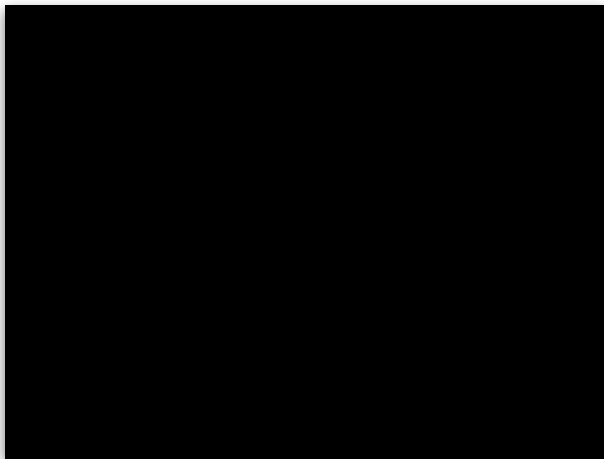
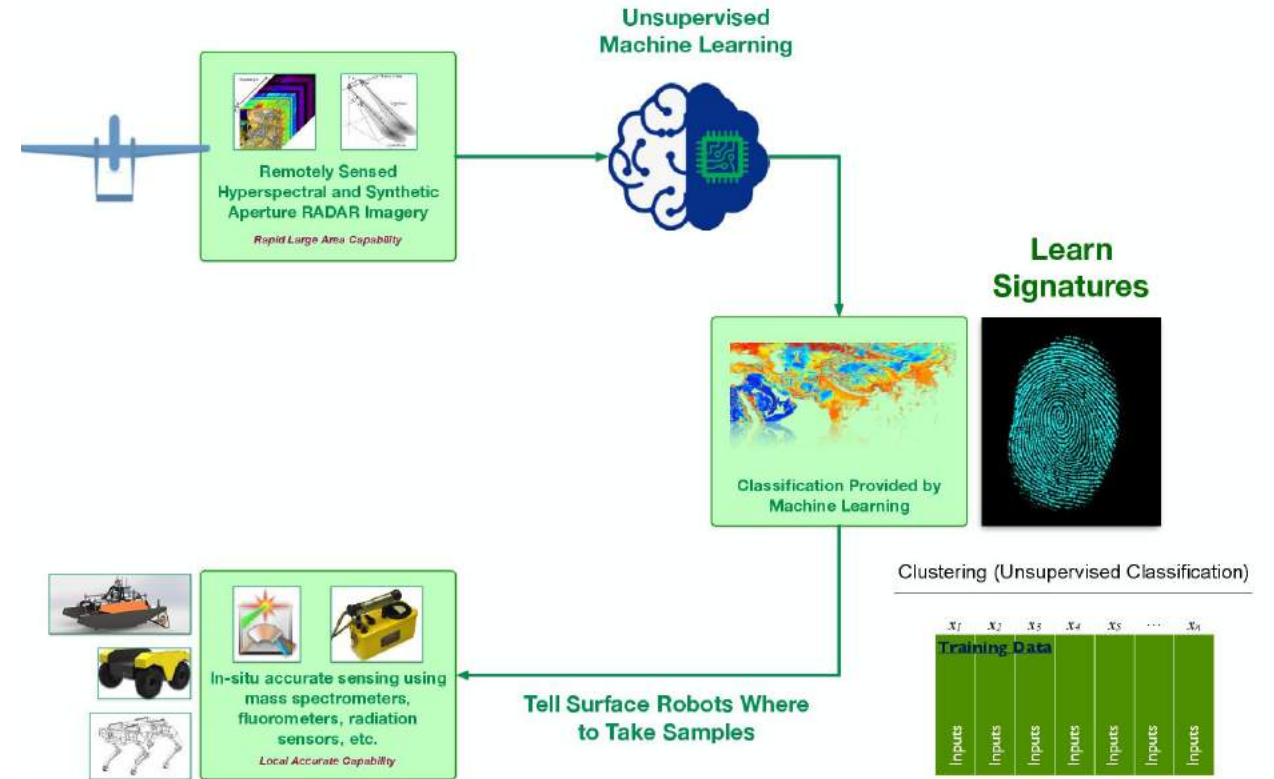
Mode 1: Coordinated robots using onboard Machine Learning for specific data products



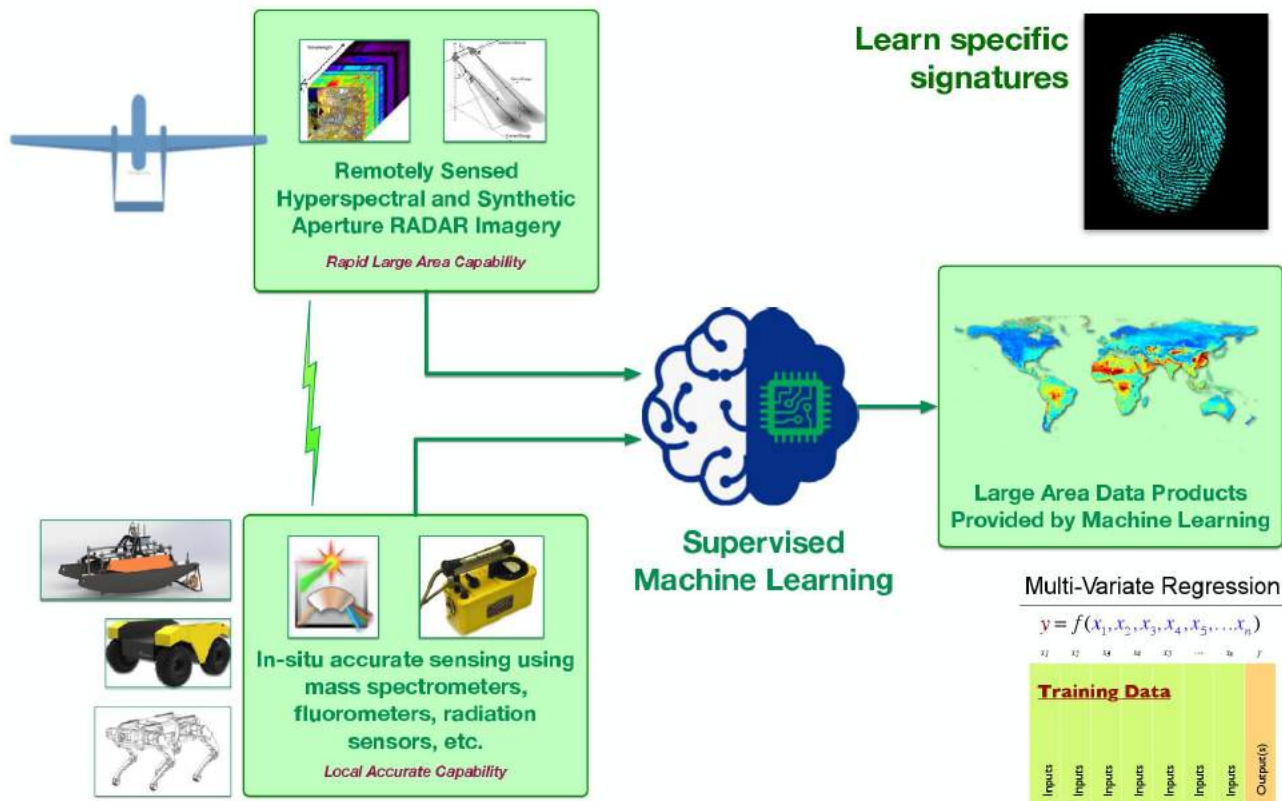
Mode 1: Coordinated robots using onboard Machine Learning for specific data products



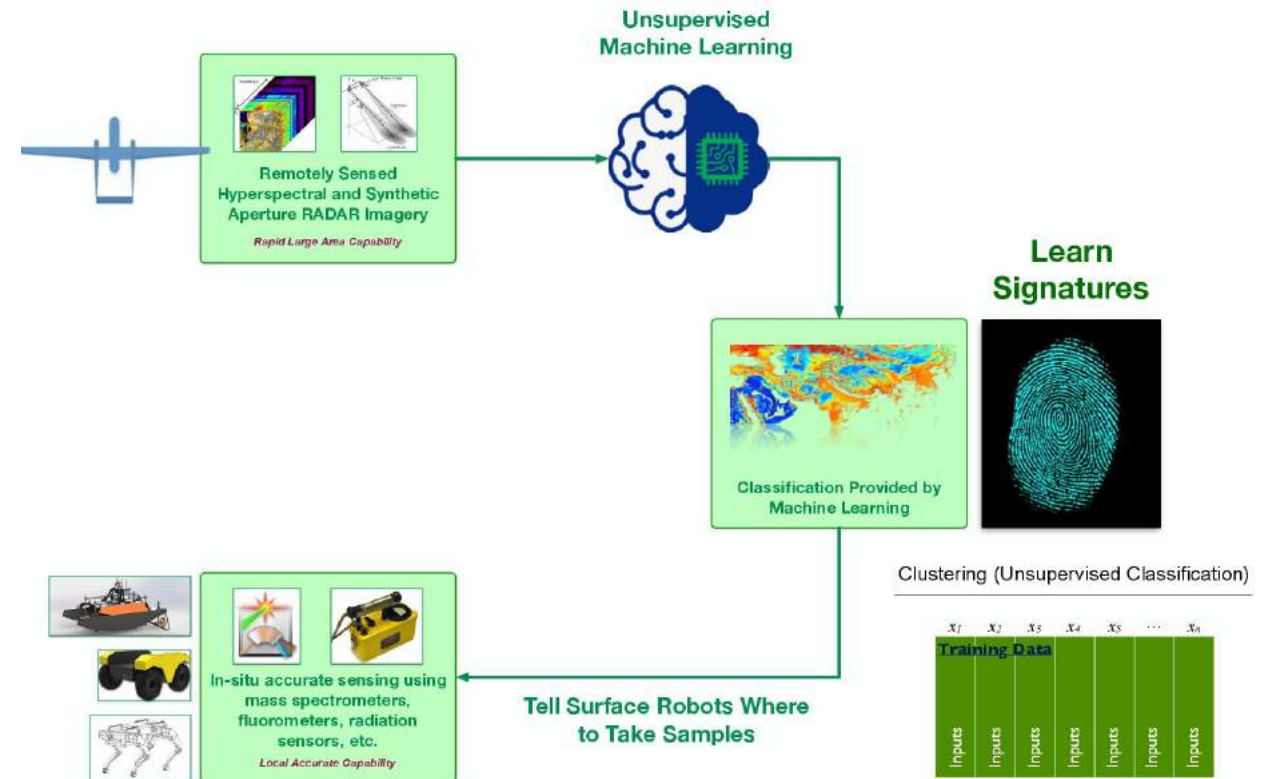
Mode 2: Unsupervised classification



Mode 1: Coordinated robots using onboard Machine Learning for specific data products



Mode 2: Unsupervised classification



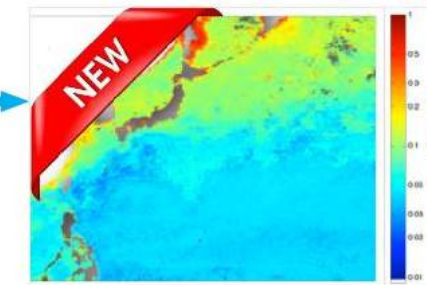
Off the shelf Aerial Vehicle



Off the shelf Hyperspectral Imaging



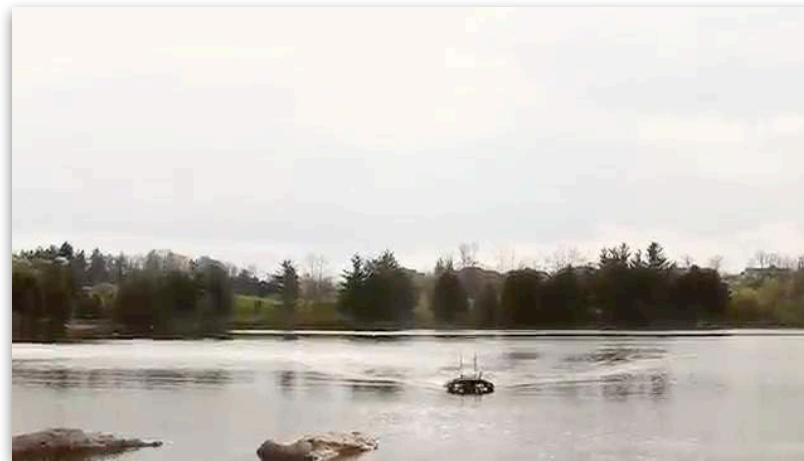
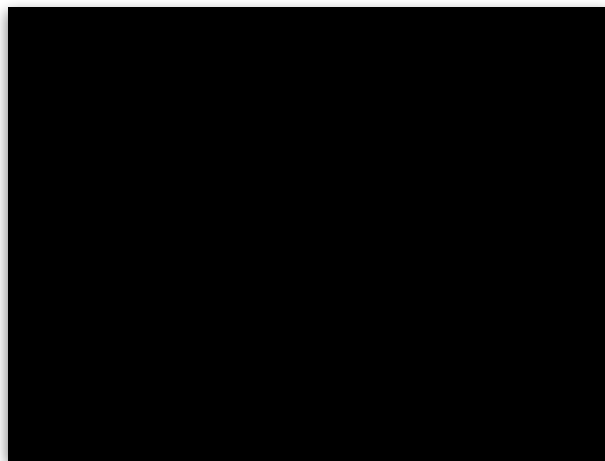
Embedded Intelligence with real-time Machine Learning



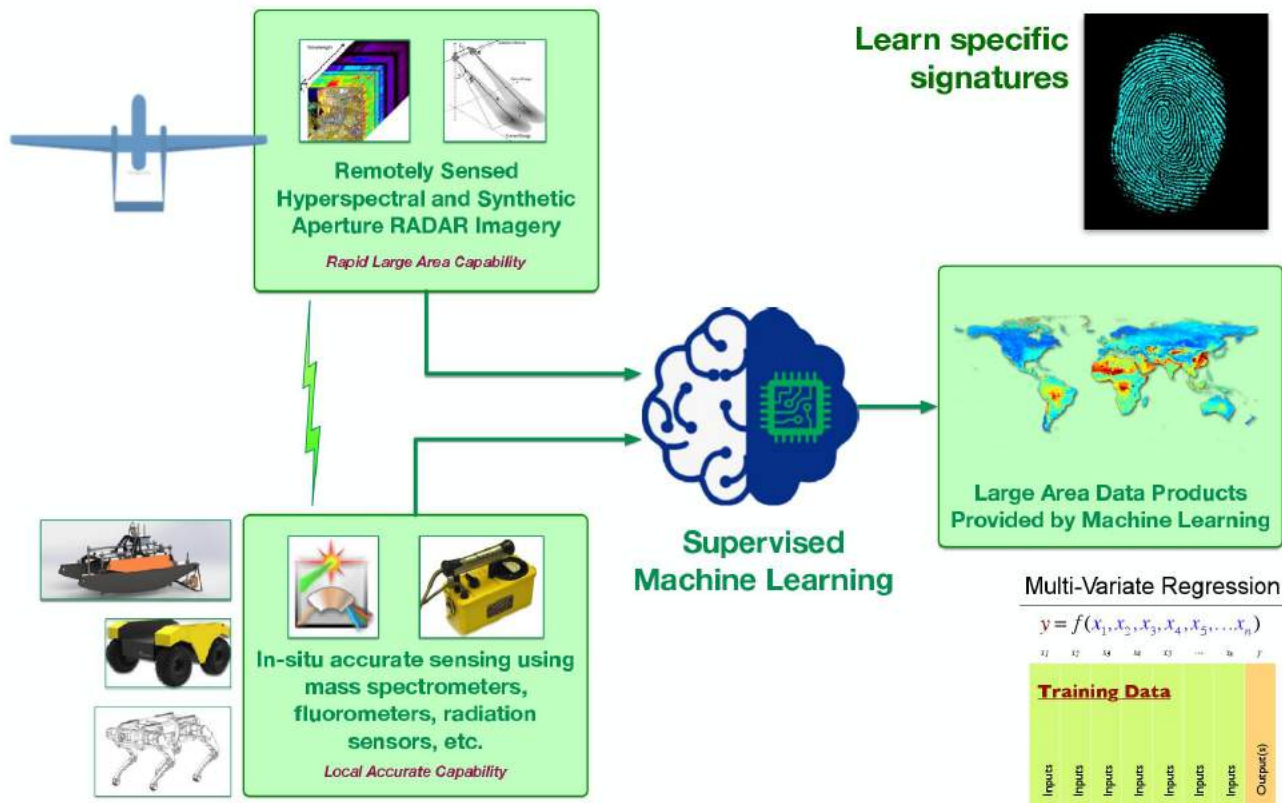
Onboard real-time data product creation



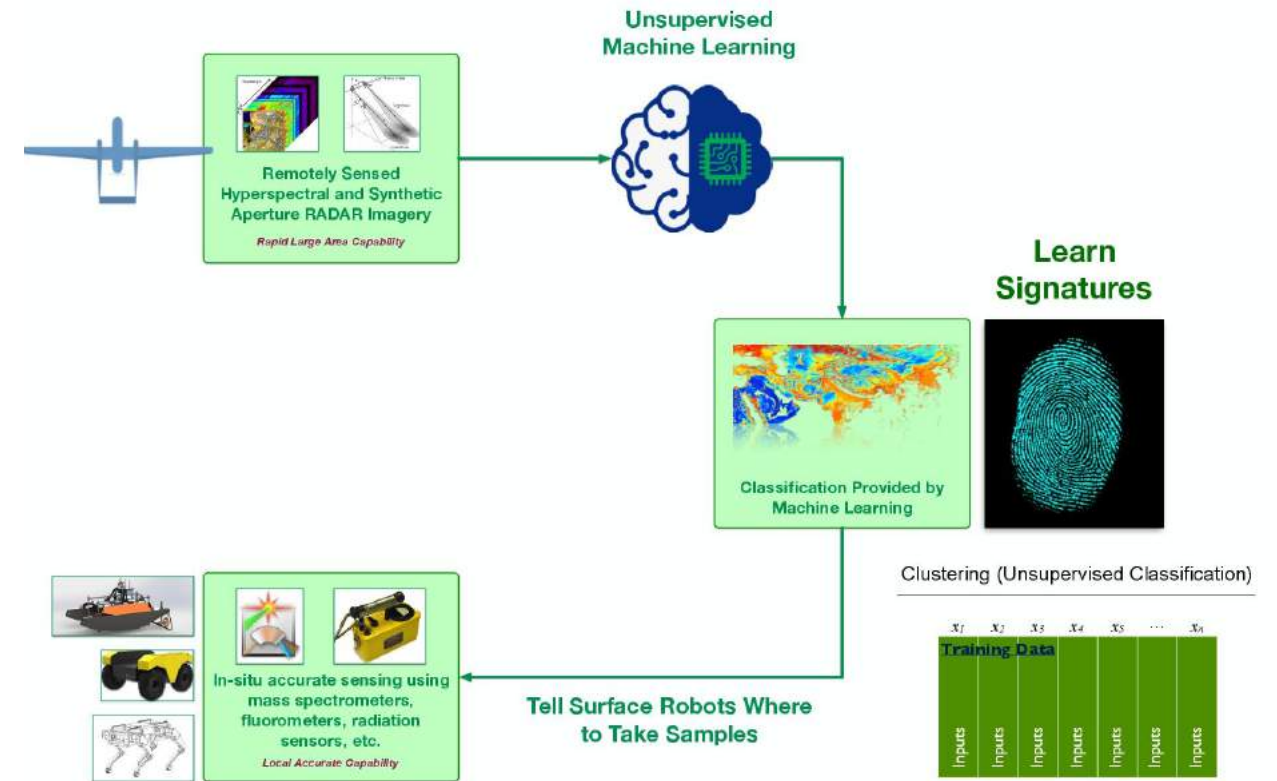
Downlink to ground station



Mode 1: Coordinated robots using onboard Machine Learning for specific data products



Mode 2: Unsupervised classification



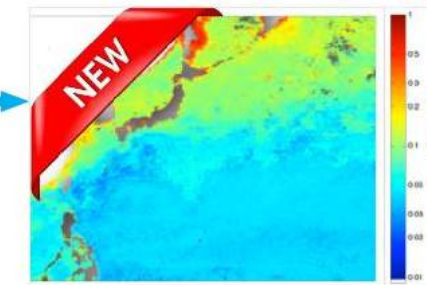
Off the shelf Aerial Vehicle



Off the shelf Hyperspectral Imaging



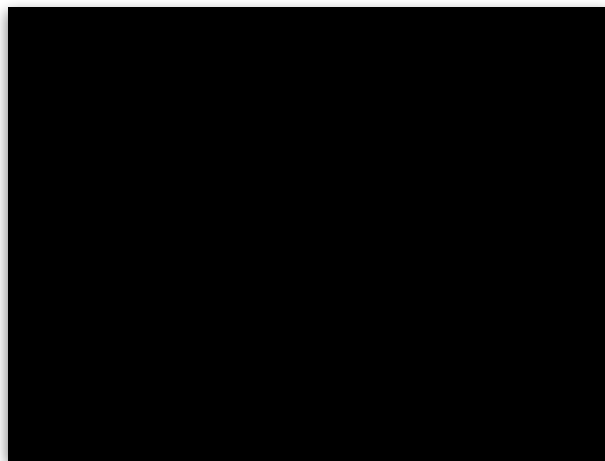
Embedded Intelligence with real-time Machine Learning



Onboard real-time data product creation



Downlink to ground station



Back Up Slides

What is Machine Learning?

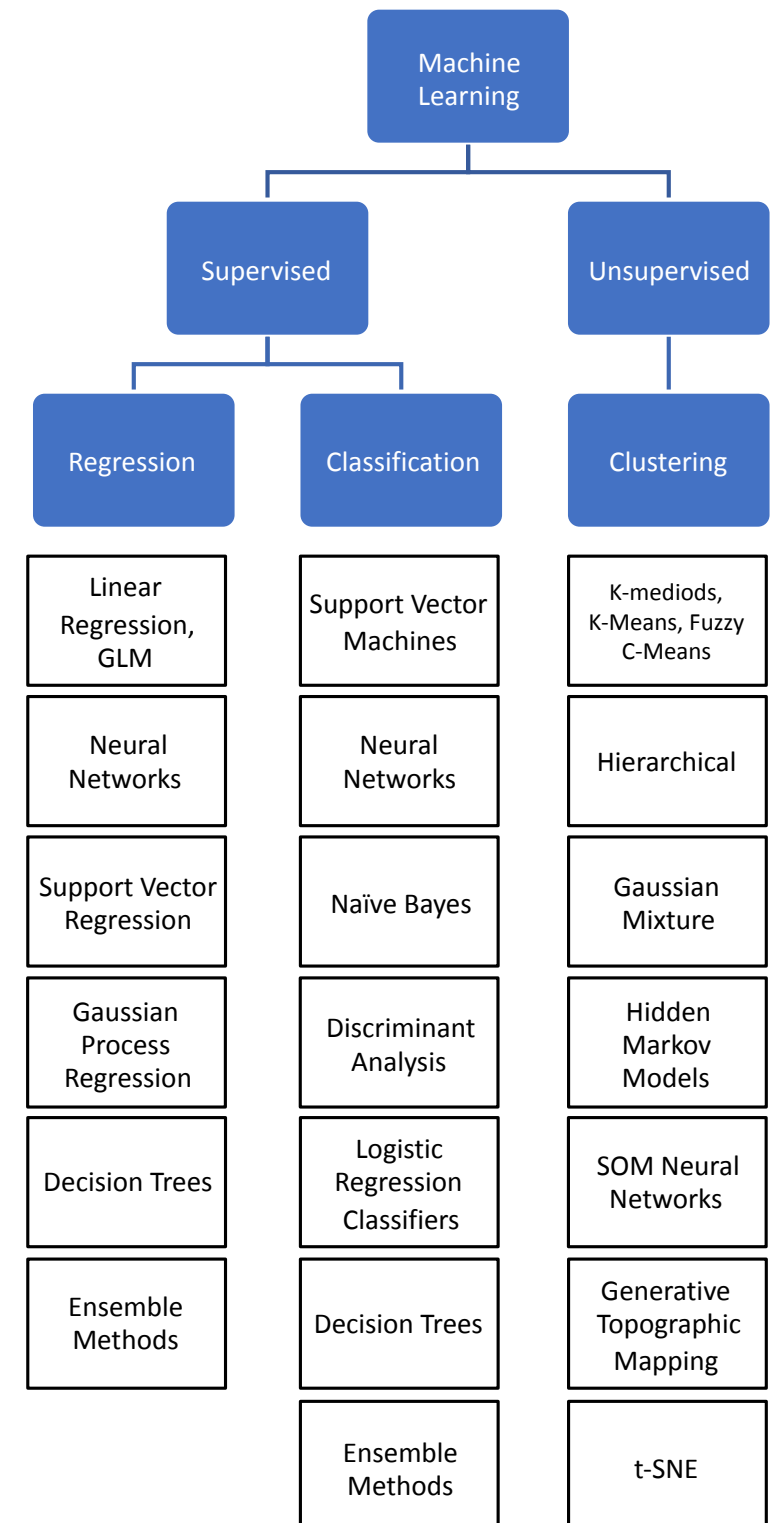
Machine learning is an **automated approach** to building **empirical models** from the **data alone**.



Arthur Lee Samuel
1901-1990

As Arthur Samuel coined the term Machine Learning in 1959. He said machine learning gives *'computers the ability to learn without being explicitly programmed.'*

Just as **humans learn by experience**, machine learning algorithms let **computers learn from data**.

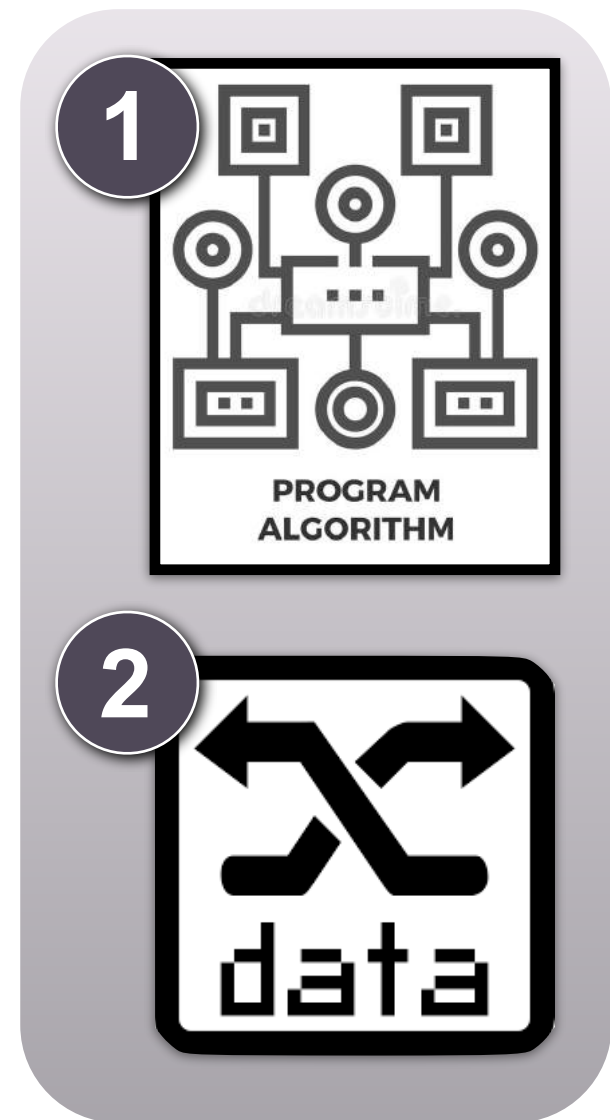


Two Key Ingredients

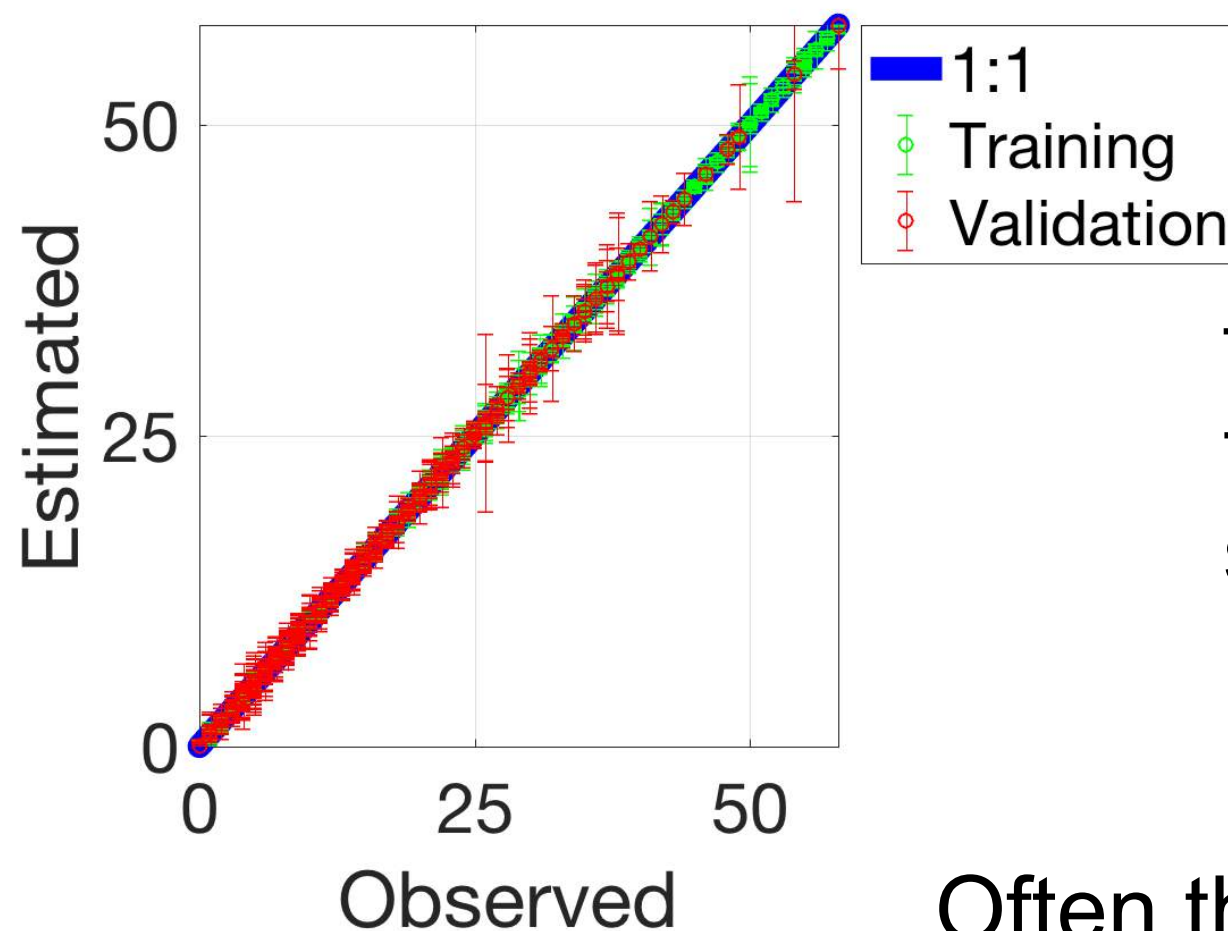
For a successful application of machine learning we have **two** key ingredients, **both** of which are **essential**:

- A machine learning algorithm,
- A comprehensive training dataset that the algorithm can learn from.

Once the training has been performed, the empirical model should always be **tested** using an **independent validation** dataset to see how well it performs when presented with data that the algorithm has not previously seen, i.e. we need to test the **generalization**. This can be, for example, a randomly selected subset of the training data that was held back and then utilized for independent validation.



Independent Verification

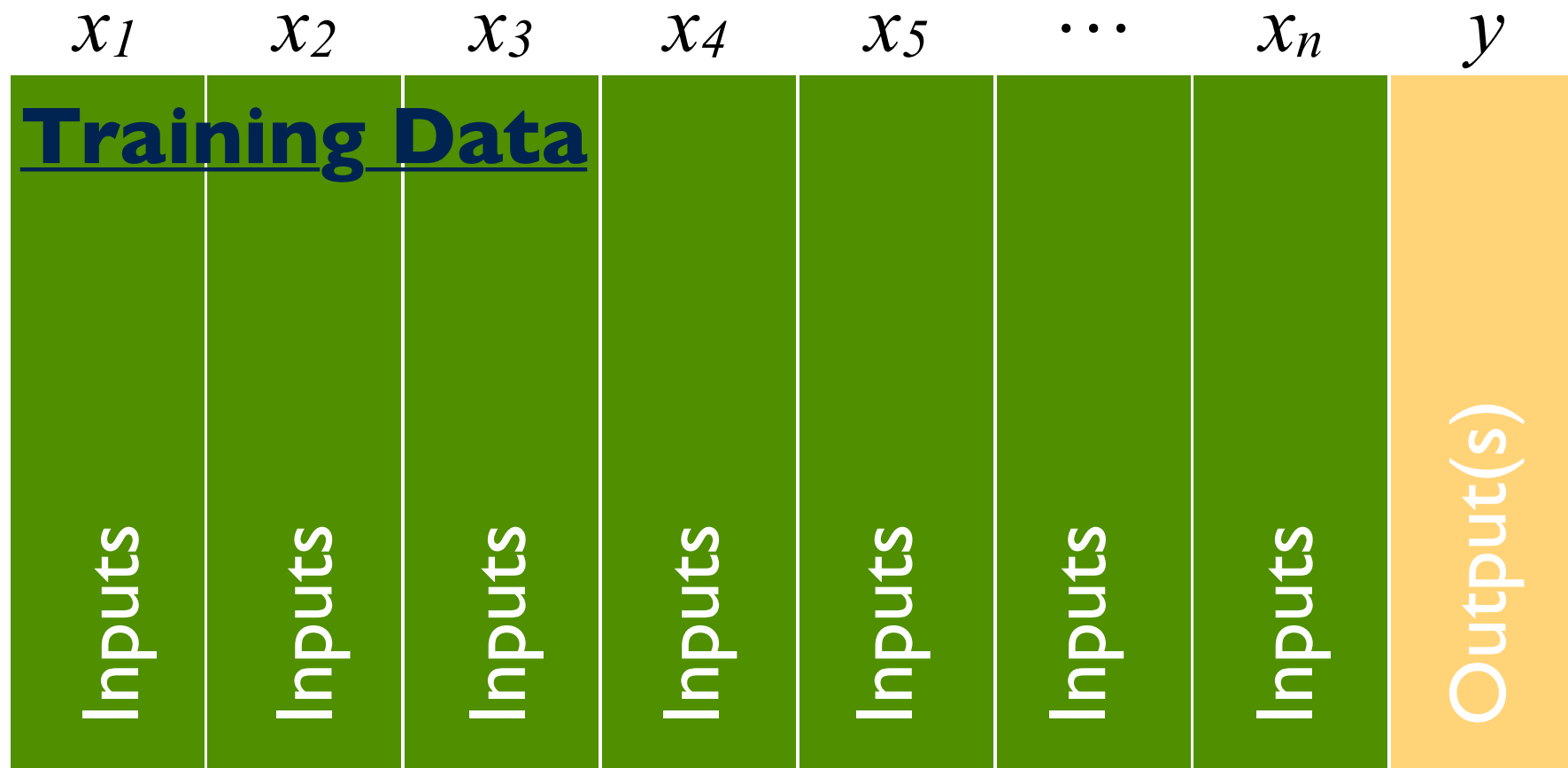


It is critical to emphasize that before using a machine learning model operationally, or for key decision making tasks, the model's performance should always be **independently verified**.

Often the training data will be randomly split up into two portions. One portion is used for training, the other is not used in the training but reserved for an independent validation once the training is complete.

Regression

$$y = f(x_1, x_2, x_3, x_4, x_5, \dots, x_n)$$



Multivariate, nonlinear, nonparametric
 n can be very large

